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

Mor Absa Loum [1*], Marie-Anne POURSAT [1], Abdourahmane SOW [2], Amadou Alpha SALL [2], Cheikh LOUCOUBAR [3], Elisabeth Gassiat [1]

[1] *Laboratoire de Mathématiques d'Orsay, Universié Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, FRANCE*

[2] *Institut Pasteur de Dakar, Arboviruses and Viral Hemorrhagic Fevers Unit*

[3] *Institut Pasteur de Dakar, Biostatistics, Bioinformatics and Modeling Group*

Abstract

In tropical regions, populations continue to suffer morbidity and mortality from malaria and arboviral diseases. In Kedougou (Senegal), these illnesses are all endemic due to the climate and its geographical position. The co-circulation of malaria parasites and arboviruses can explain the observation of coinfected cases. Indeed there is strong resemblance in symptoms between these diseases making problematic targeted medical care of coinfected cases. This is due to the fact that the origin of illness is not obviously known. Some cases could be immunized against one or the other of the pathogens, immunity typically acquired with factors like age and exposure as usual for endemic area. Then, coinfection needs to be better diagnosed. Using data collected from patients in Kedougou region, from 2009 to 2013, we adjusted a multinomial logistic model and selected relevant variables in explaining coinfection status. We observed specific sets of variables explaining each of the diseases exclusively and the coinfection. We tested the independence between arboviral and malaria infections and derived coinfection probabilities from the model fitting. In case of a coinfection probability greater than a threshold value to be calibrated on the data, duration of illness above 3 days and age above 10 years-old are mostly indicative of arboviral disease while body temperature higher than 40^oC and presence of nausea or vomiting symptoms during the rainy season are mostly indicative of malaria disease.

*Corresponding author: mor-absa.loum@u-psud.fr

Keywords: Arbovirus, coinfection, malaria, multinomial logistic regression, random forest classification, variable selection.

1. Introduction

Concurrent infections are often observed among vector borne diseases such as malaria and arthropod-borne viral diseases (arbovirus) in tropical regions ([1, 2]). It is the case for malaria and dengue in American, African and Asian tropical regions where their endemic areas overlap largely ([3, 4, 5, 6, 7, 8, 9]). Malaria can be easily ascribed to other febrile illnesses because its clinical symptoms are often indistinguishable from those initially seen in dengue or chikungunya for instance ([10]). Since the introduction of the Rapid Diagnostic Test (RDT) in 2007 in Senegal, malaria has been better diagnosed and an important decrease had been noticed in the prevalence of malaria. Thus we may think that malaria has been overestimated for some time at the expense of other febrile diseases such as arbovirus or bacteria ([11, 12]). Presumptive treatment of fever with antimalarial is widely practiced to reduce malaria attributable mortality. This practice means that ill patients may be inappropriately treated, particularly where rapid diagnosis test kits are not readily available, or if the opportunity to test for arboviral infections is missed. Thus, misdiagnosis of arbovirus coinfections as malaria infections may be a cause for underestimating emerging arbovirus infections. In 2009, surveillance of acute febrile illness (AFI) was implemented in Kedougou for early detection of arbovirus outbreaks and malaria and in order to accurately measure disease morbidity and mortality in this geographical region. Due to co-circulation of malaria parasites and arbovirus, that were mainly dengue (DEN), chikungunya (CHIK), Zika (ZIK), yellow fever (YF) and Rift Valley fever viruses (RVFV) in this region (neglecting the prevalence of other arboviral infections), concurrent infections were observed and posed a challenge for medical diagnosis ([13]). Here we compare clinical profiles of coinfected patients to clinical profiles of mono-infected patients through the statistical analysis of a data set collected from febrile patients in the Kedougou region, Senegal from 2009 to 2013. Our study aims to characterize the risk factors of coinfection and to provide statistical indicators that improve differential diagnosis of febrile cases for arbovirus.

The data of our study were provided by the Institut Pasteur de Dakar (IPD) at Kedougou (southern-east Senegal). In this region, malaria and arbovirus

are endemic due to the climate and the population movements. Data were collected through seven healthcare centers in the region: *Ninefesha rural hospita*l, *Kedougou* and *Saraya Health Centers, Bandafassi* and *Khossanto health posts*, *the Kedougou military health post,* and the *Catholic Mission mobile team*. Inclusion criteria were (i) being at least one year old at the date of the visit, (ii) having fever (i.e., body temperature$\geq 38^oC$) and (iii) manifesting at least one clinical sign within a list of symptoms. Patients satisfying inclusion criteria were enrolled once a written informed consent was signed.

In the present paper, we propose a multinomial logistic model to analyse coinfections between arbovirus and malaria. There were four outcomes determining four groups of patients: arbovirus monoinfections (with respect to the 5 tested arbovirus), malaria monoinfections, arbovirus-malaria coinfections and controls defined as patients negative for malaria and for the 5 tested arbovirus. Febrile episodes from this control group were probably due to other circulating pathogens for which all groups were supposed to be equally exposed. We performed a covariable selection using random forests based on the variable importance measure ([14]). Then we fitted a parametric multinomial logistic model including the selected covariables and we proposed a Wald-type test to test the correlation between malaria infection and arboviral infection. As the independence hypothesis was rejected, we were able to predict the probability that a patient be coinfected given that malaria is observed. From the analysis of the influent factors on the different outcomes, we investigated the following questions: Which factors can explain coinfection? Which risk factors enable to distinguish between malaria and arbovirus?

The paper is organized as follows. In Section 2, we present the working data set. Section 3 describes the statistical model and the variable selection. In Section 4, we present the independence test between arbovirus and malaria infections and we propose a predictive analysis. A concluding discussion is given in Section 5.

2. Data description

We based our analysis on the data from the Institut Pasteur de Dakar (IPD) at Kedougou. The initial data set included 15 523 patients and collected various features: patients' data (like sex, age, occupation, location,...), clinical symptoms, climate indicators and three binary infections status vari-

ables indicating (i) the presence or absence of malaria parasites in blood, (ii) detection of virus or IgM antibodies against virus and (iii) detection of IgG antibodies against virus. Malaria diagnosis relied on the identification of haematozoa using the thick blood smear (TBS) method. Arboviral infections were investigated by the detection of specific anti-arbovirus IgM and/or IgG antibodies using ELISA (enzyme-linked immunosorbent assay). We considered an *arboviral* case as any individual tested positive to the infection with at least one of the five arbovirus (DEN, CHIK, ZIK, YF and RVF).
Based on these data we created a new categorical response variable built from four possible combinations of the three infection status variables as follows:

$$Y = \begin{cases} 0 & \text{“Other febrile illnesses (O)”} \\ 1 & \text{“Arboviral monoinfection (A)”} \\ 2 & \text{“Malaria monoinfection (M)”} \\ 3 & \text{“Coinfection (C)”} \end{cases}$$

Category 0 corresponds to individuals that are negative for both malaria and the tested arboviral infections; their symptoms could be due to other unknown febrile illnesses. Category 1 corresponds to individuals positive for at least one of the five tested arbovirus and negative for malaria. Category 2 corresponds to individuals negative for tested arbovirus and positive for malaria. Category 3 represents individuals simultaneously positive for malaria and for at least one of the tested arbovirus. The subjects of category 3 are said “coinfected” with malaria and arbovirus.
Our aim is to differentiate febrile syndroms that could be due to arbovirus from febrile syndroms that could be due to malaria. As coinfection in a single patient may change the spectrum of clinical symptoms, we want to identify those features that predict arboviral infection to improve medical and treatment diagnosis in the primary care setting.

2.1. Data set

In this study, arboviral cases are diagnosed by the detection of IgM or IgG antibodies. We can have two different ways of defining an arboviral case: (1) by considering only the detection of IgM antibodies or (2) by considering the detection of both IgM and IgG antibodies. Biologically, IgM detection among patients means that they have a recent arboviral infection. So we considered that positive IgM cases are positive arboviral cases. Ignoring individuals with missing data (974 missing data on Malaria response and 803 missing data on

the covariates values), we obtained a data set of size $n = 12\ 288$, called the IgM data set. We noticed that the distributions of the different variables with and whithout missing data remain similar. A summary of the *IgM data* is given in Table 1. We can see that this data set is very imbalanced (3 arboviral or coinfected cases per 1 000 patients) and it will require a specific statistical analysis.

Malaria / **Arbovirus**	+	−	Total
+	18 (0.15%)	21 (0.16%)	39 ($A+$)
−	7 069 (57.53%)	5 180 (42.16%)	12 305 ($A-$)
Total	7 087 ($M+$)	5 201 ($M-$)	12 288

Table 1: IgM data. $A+$ for the individuals positive to arboviral infection, $A-$ for the individuals negative to arboviral infection, $M+$ for the individuals positive to Malaria and $M-$ for the individuals negative to Malaria.

The diagnosis of arboviral infection at the time of an acute episode is ideally based on the presence in the serum of a patient of detectable IgM. However, to obtain a more balanced data set, we decided to build a separate data set by considering arboviral infected patients as individuals who were tested positive to IgM or IgG. 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 2. 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.

Thereafter, we will consider two data sets that are derived from the same original data set using two different encoding: 1. the IgM/IgG data set which is suitable to apply our entire methodology; 2. the IgM data set containing the true arboviral status (from a biological point of view) which is strongly imbalanced. We will use in the next section a re-sampling strategy to handle this problem.

Malaria / **Arbovirus**	+	−	Total
+	397 (20.10%)	263 (13.31%)	633 ($A+$)
−	751 (38.00%)	565 (28.59%)	1318 ($A-$)
Total	1148 ($M+$)	828 ($M-$)	1976

Table 2: IgM/IgG data. $A+$ for the individuals positive to arboviral infection, $A-$ for the individuals negative to arboviral infection, $M+$ for the individuals positive to Malaria and $M-$ for the individuals negative to Malaria.

2.2. Covariates

In this data set, there are four quantitative covariables: the measured body temperature (in Celsius degrees), the number of sick days defined as the number of days between the date of symptoms onset and the date of consultation, the patient's age (in year) and the rainfall measure (in millimiters) which is a proxy for the season (rainy or dry). The individual rainfall measure corresponds to the rainfall measure of the patient's month of consultation. The eleven qualitative covariables are the patient's gender and ten other binary variable, which record presence or absence of ten symptoms: headache, eye pain, muscle pain, join pain, cough, nausea or vomiting, chills, diarrhea, nasal congestion and icterus and/or jaundice. All the variables of the data sets are summarized in Figure 1.

In our data set, females represented 42% of the population and males represented 58% of the population. In the IgM data set, the two categories "Coinfection" and "Arboviral monoinfection" are underrepresented, which results in irrelevant descriptive graphs. 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 coinfected. Among the patients having a nasal congestion symptom, 31% were positive to malaria monoinfection, 21% were coinfected and 14% were positive to arboviral monoinfection. Figure 2 displays the distributions of age, rainfall and number of sick days over the four classes of the IgM/IgG data set. Overall, Figure 2 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 2(b) shows that high values of rainfall are

	For categorical variables			quantitative variables			
Designation	# levels	0 (%)	1 (%)	mean	median	min	max
Age				19.5	16.5	1	90
Temperature				38.97	39	38	42
Number of sick days				3.039	3	0	19
Rainfall				147.5	76.1	0	500.2
Sex (F=0 and H=1)	2	42	58				
Cephalalgia	2	6	94				
Nausea/vomiting	2	50	50				
Diarrhea	2	83	17				
Chills	2	45	55				
Cough	2	64	36				
Eye pain	2	95	5				
Joint pain	2	77	23				
Muscl pain	2	71	29				
Nasal congestion	2	54	46				
Ictere/jaudice	2	95	5				
Malaria	2	42	58				
IgM	2	99	1				
IgG	2	95	5				

Figure 1: List of variables

observed in the coinfection and malaria groups.

3. Statistical analysis of the coinfection influential factors

The objective of this section is to propose a methodology that can identify the important symptoms for the arbovirus diagnosis and can help making decision for arbovirus treatment in absence of laboratory confirmation.
Variable selection is appreciable in medical data analysis as the diagnosis of the disease could be done on a minimum number of clinical measures. Reducing the number of relevant covariates may also produce more accurate classification results. In a first step, we select relevant covariates that explain the disease status typically via a multinomial logistic model. The statistical analysis is challenging because of the small number of instances of the arboviral class (39) with respect to the total number of observations (12 288). The cases that are more important for the study are rare and few exist on

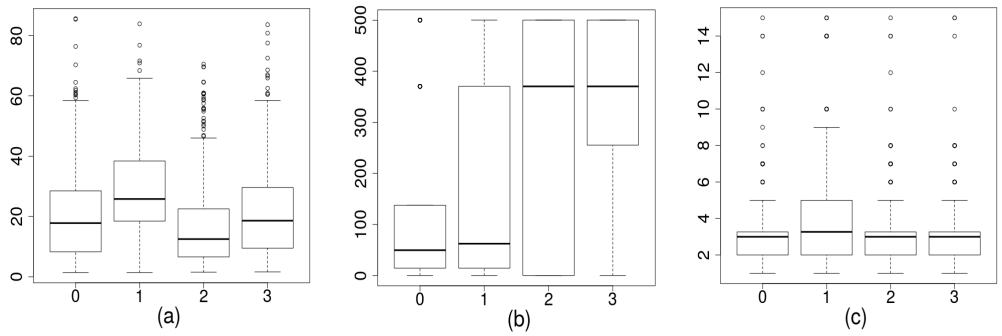

Figure 2: *IgM/IgG* data set; boxplots of the empirical distributions of the covariates (a) *age*, (b) *rainfall* and (c) *number of sick days* for the four modalities of the response variable Y : 0 (other febrile illnesses), 1 (arboviral monoinfection), 2 (malaria monoinfection) and 3 (coinfection).

the available training set. We face what is usually known as a problem of imbalanced data sets. This results in models that poorly represent the rare class examples or simply ignore the observations of the minority class. To handle this problem, we proposed two data pre-processing approaches. The first one is based on biological considerations and extends the arboviral cases from 39 to 633 by merging patients with either blood positive IgM or blood positive IgG. We obtained the balanced IgM/IgG data set described in the previous section, which contains 1976 observations. The second approach involves randomly removing observations from the majority class to prevent its signal from dominating the fitting procedure. We applied to the imbalanced IgM data set a common undersampling technique to obtain a more balanced data distribution. As the data distribution is changed, it is expected that the fitted models are biased to the goals of the user and are more interpretable in terms of these goals.

In a second step we investigate the robustness of the variable selection using random forests. Introduced by [15], random forests (RF hereafter) are a robust nonparametric method to deal with classification problems. They require only mild conditions on the data generating model. They are also less sensitive to weaknesses in the data, because the randomized tree generation procedure ensures that all covariates are more equally evaluated. Moreover, RF decision trees often perform well on imbalanced data sets because ensemble methods offer ways to rebalance the distributions in varied ways. In this

study, RF models have the advantage of providing a ranking of covariates using the RF score of variable importance that is a useful and effective tool to find important covariates for interpretation.
In a third step, we quantify the effects of the selected covariates using odds ratios. We compute odds ratios for one disease category relative to an other one and we contrast the effects of the covariates on the disease category, arboviral monoinfection, malaria monoinfection and coinfection.

3.1. Multinomial logit model

We recall that Y is the response variable indicating the class of the disease: "Other febrile illnesses" ($Y = 0$), "arboviral monoinfection" ($Y = 1$), "malaria monoinfection" ($Y = 2$) and "coinfection" ($Y = 3$). Let $X = (1, X_1, \ldots, X_p)$ be the vector of the p covariates. For an individual with covariates $X = x$, we want to predict the probability of belonging to the class k given x,

$$\pi_k(x) = \mathrm{P}(Y = k|X = x), \quad k = 0, 1, 2, 3.$$

The multinomial logit model assumes the existence of β_1, β_2, $\beta_3 \in \mathbb{R}^{p+1}$ such that, for each $k = 1, 2, 3$ and each vector of covariates x,

$$\log \frac{\mathrm{P}(Y = k|X = x)}{\mathrm{P}(Y = 0|X = x)} \quad = \quad \langle x, \beta_k \rangle \tag{1}$$

where

$$\langle x, \beta_k \rangle = \sum_{j=0}^{p} x_j \beta_{kj}$$

and $x_0 = 1$ to include the intercept parameters β_{k0}, $k = 1, 2, 3$. The reference modality is class 0.
Consequently, for each $k = 1, 2, 3$ and each vector of covariates x,

$$\mathrm{P}(Y = k|X = x) = \frac{\exp(\langle x, \beta_k \rangle)}{1 + \sum_{l=1}^{3} \exp(\langle x, \beta_l \rangle)}$$

and

$$\mathrm{P}(Y = 0|X = x) = \frac{1}{1 + \sum_{l=1}^{3} \exp(\langle x, \beta_l \rangle)}.$$

From the computation of the maximum likelihood estimates $\widehat{\beta}_k$, we derive for $k = 1, 2, 3$,

$$\widehat{\pi}_k(x) = \frac{e^{\langle x, \widehat{\beta_k} \rangle}}{1 + \sum_{l=1}^{3} e^{\langle x, \widehat{\beta}_l \rangle}}. \tag{2}$$

3.1.1. Application to the IgM/IgG data

We first give the results for the IgM/IgG data set since they are based on a standard logit analysis. 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 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 $7.44\ 10^{-3}$.

3.1.2. Fitting strategy for handling imbalanced IgM data

The IgM data set contains 18 arboviral monoinfection cases, 21 coinfection cases, 5 180 other febrile illness cases and 7 069 malaria monoinfection cases. Trained on the original IgM data set, the fitted logit model only predicted classes 0 and 2, which means it ignores the two minority classes 1 and 3 in favour of the majority classes.

Applying resampling strategies to obtain a more balanced data sample is an effective solution to the imbalance problem (see [16] for a survey of existing methods). Two of the most simple resampling approaches are undersampling and oversampling. Since the IgM is highly imbalanced with a large number of observations in the two majority classes, we used a random undersampling strategy that removes observations and reduces the sample size. We sampled without replacement 50 cases from each of the two majority classes to create a balanced sub-sample of size $18+21+50+50 = 139$. Trained on a sub-sample, the model predicted four classes.

Undersampling results in loss of information and the risk of removing relevant observations is present. To overcome this problem, we repeated the sampling step a thousand times and worked with 1 000 balanced sub-samples of the IgM data set. The multinomial model was fitted to each sub-sample and a stepwise covariate selection was performed. The observed variability of

the 1 000 covariate selections raised robustness questions. To answer this point, we conducted a nonparametric analysis based on the RF algorithm. In recent years, several methods involving the combination of resampling and ensemble learning have appeared in the imbalanced distributions literature ([16]). We found that the importance score based on random forests yielded a convenient way to summarize the information obtained from the 1 000 sub-samples.

3.2. Variable selection using random forests

A random forest is an ensemble of unpruned trees, induced from bootstrap samples of the training data, that uses random covariate selection in the tree construction process. Prediction is made by aggregating the predictions of the ensemble, using the majority vote rule.

One of the most widely used RF score of importance of a given variable is the Mean Decrease of Accuracy (MDA) in predictions. It is based on the out-of-bag (OOB) error. For each tree t of the forest, consider the associated OOB_t sample (data not included in the bootstrap sample used to construct t). Denote by $errOOB_t$ the misclassification rate of tree t computed on this OOB_t sample. Then, randomly permute the observed values of covariate X_j in OOB_t to get a perturbed sample and compute $errOOB_t^j$, the error of t on the perturbed sample. Variable importance of X_j is then given by

$$MDA(X_j) = \frac{1}{ntree} \sum_{t=1}^{ntree} \left(errOOB_t^j - errOOB_t \right),$$

where $ntree$ denotes the number of trees of the RF. The higher the MDA, the more important the variable is. Several variable selection procedures using RF are based on this quantification of variable importance.

Using R packages, we made the following implementation choices: `randomForest` for RF fitting and MDA calculation, `VSURF` for selecting the important variables. The main parameters of `randomForest` were calibrated and set to their default values, `ntree`=500 and `mtry`=$\sqrt{p}$=3 (number of variables tried at each split of a tree of the RF). The variable selection strategy of `VSURF` is based on a two-stage procedure ([17]): 1. the covariates are ranked by sorting their variable importance measures in descending order and the covariates whose importance is less than a threshold (the minimum value of the standard deviations of the importance measures) are eliminated; 2. a sequence of nested models starting from the one with only the most important variable

and ending with the one involving all important variables kept previously is considered; the variables of the model leading to the smallest OOB error are selected. An advantage of using VSURF is that this procedure does not require the choice of tuning parameters.

3.2.1. Application to the IgM/IgG data

A graphical representation of the variable importance of the 15 covariates is shown in Figure 3. 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*.

3.2.2. Application to the IgM data

Figure 4 ranks the variable importances (MDA) of the 15 covariates across the 1 000 sub-samples. First, *rainfall* is the most important covariate; a second group of less important covariates is formed by *cough, age and joint pain*; then comes a group of five covariates: *number of sick days, temperature, nausea or vomiting, eye pain and nasal congestion*; finally, six unimportant covariates are displayed: *muscle pain, chills, cephalalgia, jaudice, diarrhea and sex* . The boundary between the two last groups is not clear and we used the VSURF procedure to separate the important covariates from the other ones. We can notice on the plot that both MDA level and variability are larger for relevant variables; as explained by [14], this is expected and the VSURF threshold value is based on MDA standard deviation estimation. Figure 5 summarizes the results of the VSURF selection procedure based on the 1 000 sub-samples. The covariate *rainfall* (95.2%) is almost always selected. Next, the more often selected variables are *cough* (29.1%), *age* (28.3%), *joint pain* (19.8%), *nausea or vomiting* (16.4%), *number of sick days* (16.1%), *temperature* (16.1%) and *nasal congestion* (11%), in decreasing order. The other covariates are selected in less than 10% of the samples.
We set different random seeds and we found that, for our purpose of selecting significant covariates, aggregation of 1 000 RF classifiers learned from 1 000

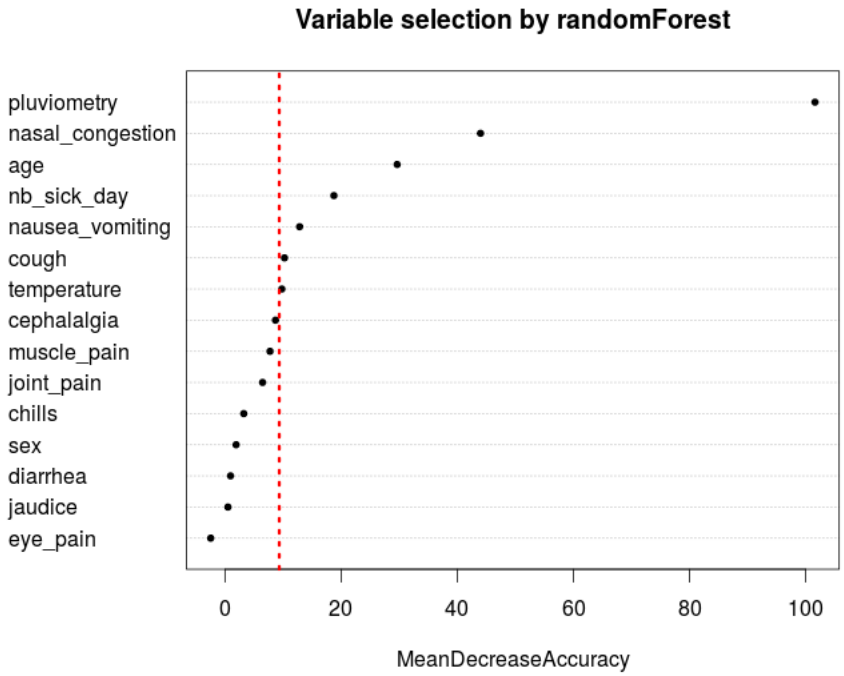

Figure 3: A variable importance plot for the *IgM/IgG* data set : mean decrease of accuracy (MDA) of the covariates, by increasing order. The variables whose MDA is to the right of the dotted line are selected by the VSURF procedure.

randomly balanced sub-samples yielded stable selected variable sets.

3.3. Influence of selected covariates on disease status

In the previous sections, we carried out a comparison between RF and multinomial logit covariate selections on the IgM/IgG data set and the conclusion is that the results are in agreement. The RF variable importance results on the IgM sub-samples produced a robust ranking of the covariates. The same group of seven important variables was selected by RF algorithm (see Figures 4 and 5); an eighth supplementary variable, *joint pain*, was added in the stepwise selection of the IgM/IgG data set. In conclusion, we decided to fit the same multinomial model with eight covariates to the data sets of our analysis and to further quantify the effects of the covariates in this model.
Within the multinomial logit model, we can quantify the effect of a variable in terms of an odds ratio or its logarithm. The odds that $Y = k$ occurs for

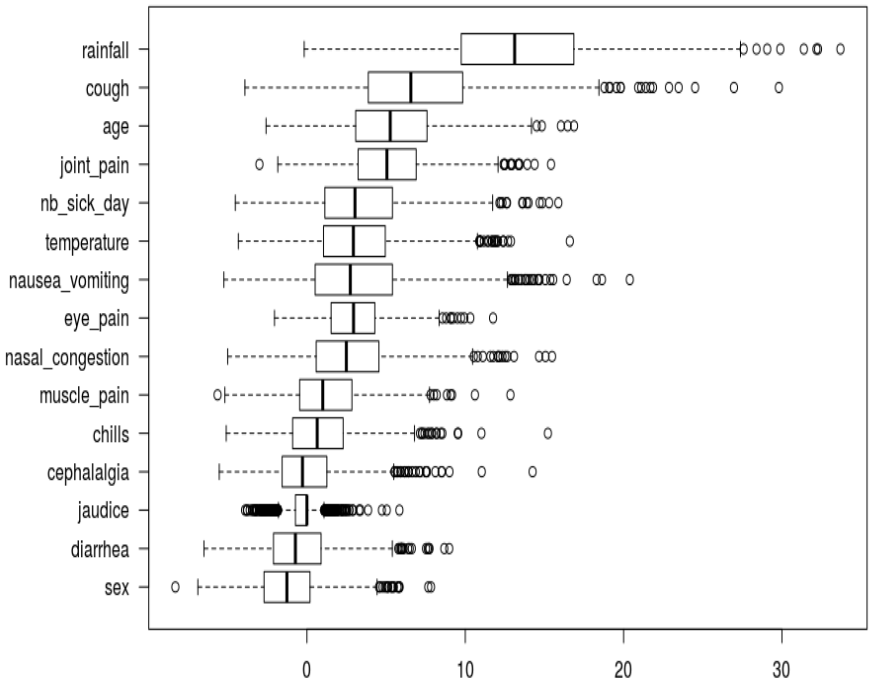

Figure 4: A variable importance plot for the *IgM* data set. Each boxplot summarizes the distribution of the variable importance among 1000 *IgM* sub-samples.

an individual with covariates $X = x$ is the ratio of $\mathrm{P}(Y = k|X = x)$ divided by $\mathrm{P}(Y = 0|X = x)$, $k = 1, 2, 3$. Then, the log odds of category k is given by Equation (1) :

$$\log \mathrm{odds}(Y = k|X = x) = \langle x, \beta_k \rangle.$$

Thus the multinomial logit model is a linear regression model in the log odds. The parameter component β_{kj} can be interpreted as the change in the log odds per unit change in the continuous covariate X_j, if all other covariates are held constant. The odds ratio (OR) of category k for a d units increase of X_j, all other covariates remaining constant, is defined as

$$OR_k(d) = \frac{\mathrm{P}(Y = k|X_j + d)/\mathrm{P}(Y = 0|X_j + d)}{\mathrm{P}(Y = k|X_j)/\mathrm{P}(Y = 0|X_j)} = \exp(\beta_{kj} d).$$

Once β is estimated, one can estimate any odds or odds ratios. An OR equal to one means that a change in covariate X_j has no effect on the odds of

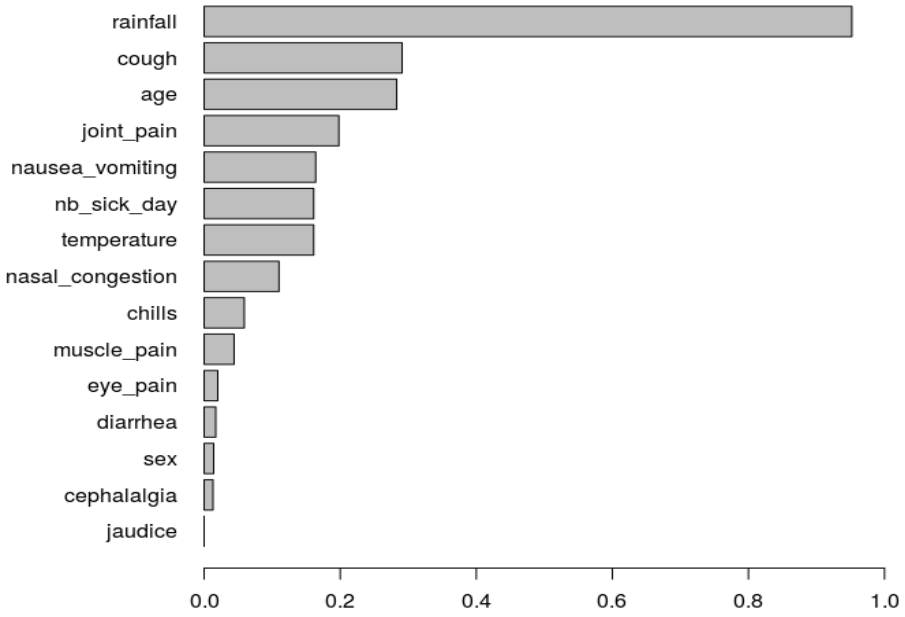

Figure 5: Ranking by VSURF: for each variable, the length of the bar corresponds to the empirical probability to be selected by VSURF among 1000 *IgM* sub-samples

category k; if $OR_k(d) > 1$ ($OR_k(d) < 1$), the effect of an increase of X_j is to increase (decrease) the odds of category k. An odds ratio is a popular description of an effect in a probability model since it can be constant. On the contrary, the risk ratio $\mathrm{P}(Y = k|X_j + d)/\mathrm{P}(Y = k|X_j)$, which could be more interpretable in terms of predicted probabilities instead of odds, depends on the values of all other covariates. ORs are similar to risk ratios if the risk is small, otherwise ORs overestimate risk ratios.

For each covariate, we computed the odds ratios OR_k, $k = 1, 2, 3$ and their confidence intervals for each disease. Table 3 for the IgM/IgG data set and Figure 7 for the IgM data set display the OR by which the odds increases for a certain change in a covariate, holding all other covariates constant. The ORs associated with binary variables (*Nausea/vomiting, Cough, Nasal congestion and Joint pain*) were computed by comparing the two modalities: 0 for absence and 1 for presence of the symptom. We computed the ORs

resulting from increasing *Temperature* from 38 to 40 degrees Celsius ($d = 2$) and from increasing *Number of sick days* from 2 to 6 days ($d = 4$). The outer quartiles of *Age* are 8 and 28 years ($d = 20$), so we computed the half-sample OR for age. Similarly, we computed the half-sample OR for a *rainfall* of 14 mm compared to a *rainfall* of 370 mm ($d = 356$).
The ORs defined previously are relative to the reference category $Y = 0$. We also computed the ORs between two diseases $Y = k$ and $Y = l$ in order to differentiate the effect of each covariable between the three clinical groups, arbovirus vs malaria, coinfection vs arbovirus and coinfection vs malaria:

$$OR_{k|l}(d) = \frac{\mathrm{P}(Y = k|X_j + d)/\mathrm{P}(Y = l|X_j + d)}{\mathrm{P}(Y = k|X_j)/\mathrm{P}(Y = l|X_j)} = \exp((\beta_{kj} - \beta_{lj})d).$$

These results are displayed in Figure 6 (IgM/IgG data set) and Figure 8 (IgM data set). The confidence intervals are derived from the fitted multinomial logit model and their accuracy is based on the parametric assumption that the true data generating distribution does fall in the model.

3.3.1. Results for the IgM/IgG data

Diseases / Variables	**Arbovirus**	**Coinfection**	**Malaria**
Age	**1.71** [1.42; 2.07]	1.12 [0.92; 1.36]	**0.61** [0.50; 0.73]
Temperature	1.02 [0.69; 1.49]	**2.16** [1.52; 3.07]	**2.47** [1.82; 3.35]
Number-of-sick-days	**2.54** [1.91; 3.37]	**1.43** [1.04; 1.96]	1.04 [0.77; 1.39]
Rainfall	**2.19** [1.53; 3.14]	**17.0** [12.0; 24.0]	**9.81** [7.18; 13.4]
Nausea /vomiting	0.83 [0.60; 1.13]	**2.07** [1.55; 2.78]	**2.15** [1.67; 2.77]
Cough	0.79 [0.58; 1.10]	**0.46** [0.33; 0.63]	**0.57** [0.44; 0.74]
Nasal congestion	**0.52** [0.35; 0.75]	**0.13** [0.09; 0.20]	**0.10** [0.07; 0.13]
Joint pain	**1.52** [0.99; 2.32]	**1.90** [1.26; 2.83]	**1.74** [1.21; 2.50]

Table 3: IgM/IgG data: odds ratios with respect to the reference modality and 95% confidence intervals.

From Table 3, 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 decreases 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.

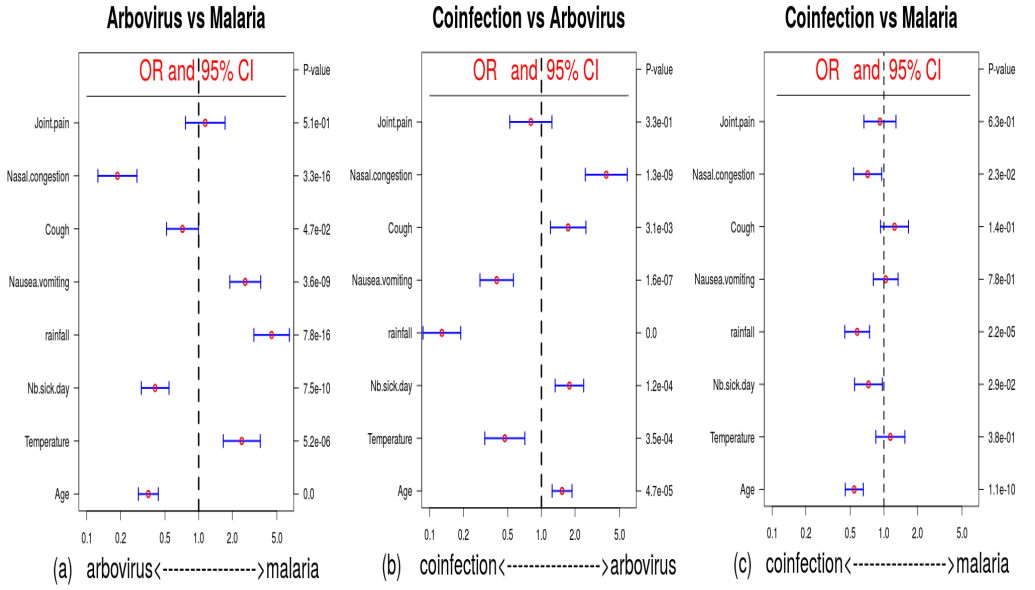

Figure 6: IgM/IgG data: odds ratios between two diseases and 95% confidence intervals; (a) Arbovirus vs Malaria (b) Coinfection vs Arbovirus (c) Coinfection vs Malaria.

Figure 6(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 vomiting* are correlated to malaria monoinfections. The variables *joint pain* and *cough* are not significant in distinguishing malaria and arboviral monoinfections. Figure 6(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 6(c)). Figure 6(c) suggests that *Age*, *Number of sick days* and *Nasal congestion* are significantly correlated

with coinfected patients compared to patients with single malaria disease.

3.3.2. Results for the IgM data

Figure 7 and Figure 8 display the sampling distribution of ORs based on the fitting of the 1000 sub-samples of the IgM data set.

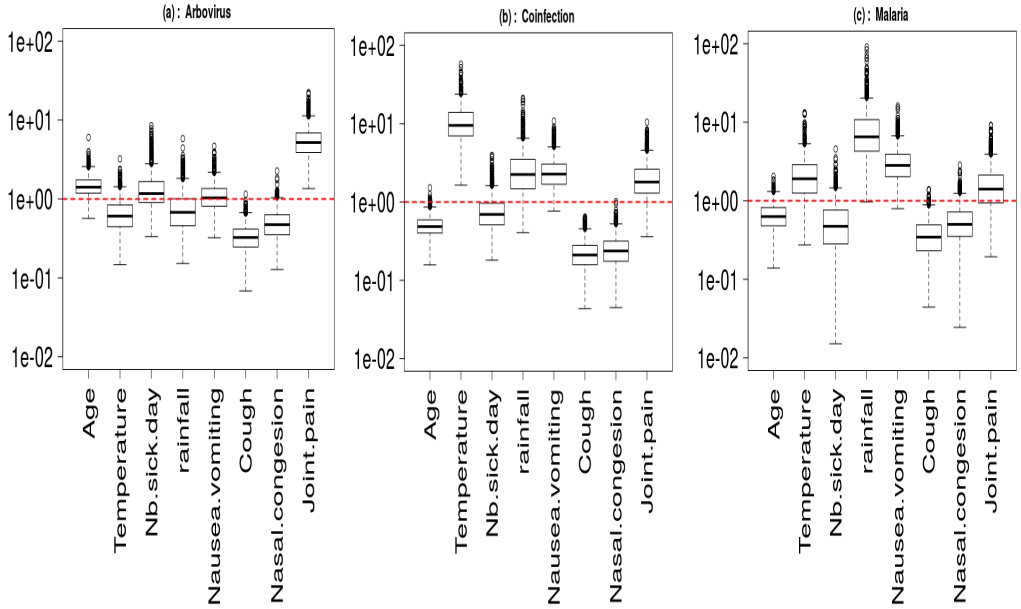

Figure 7: IgM data: boxplots of 1000 odds ratios with respect to the reference category; (a) Arbovirus (b) Coinfection (c) Malaria.

From Figure 7 and Figure 8, we can say that temperature, rainfall and vomiting symptoms are significantly correlated with malaria monoinfections whereas joint pain, age and number of sick days are correlated with arboviral monoinfections. The odds of coinfection increases with high fever and high rainfall values, and the presence of vomiting and joint pain symptoms.

3.3.3. Conclusion

The results based on both data sets show that a high temperature and the presence of nausea or vomiting symptoms are mostly indicative to malaria parasite infections whereas an increase of the number of sick days and the age are indicative to arboviral infections. The effects of the nasal congestion and joint pain symptoms on the disease status are not clear enough to be interpreted. The main question of the study was to identify risk factors that

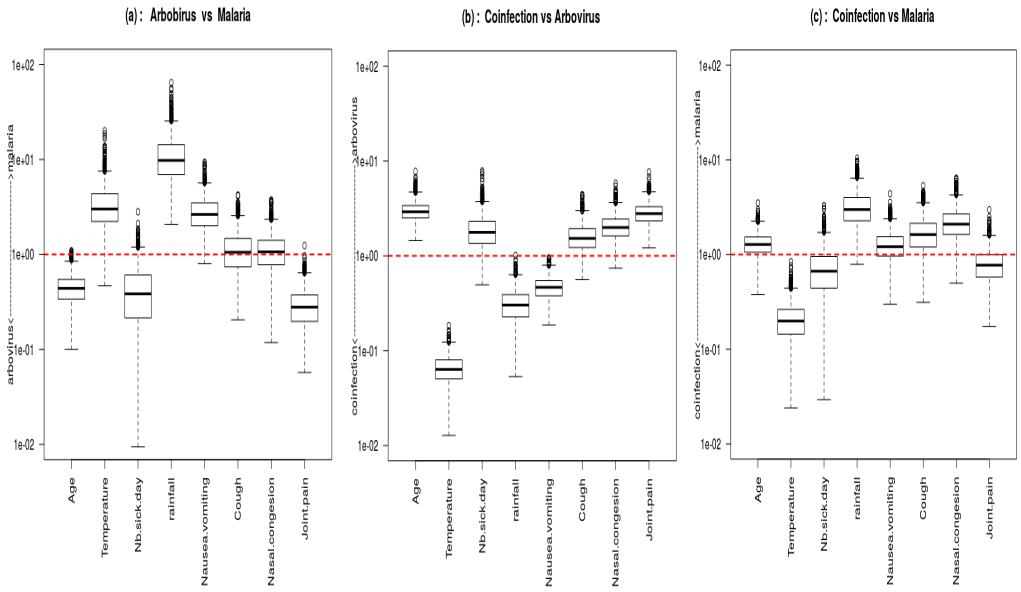

Figure 8: *IgM* data: boxplots of 1000 odds ratios between two categories; (a) Arbovirus *vs* Malaria (b) Coinfection *vs* Arbovirus (c) Coinfection *vs* Malaria.

can help doctors to diagnose a concurrent malaria and arbovirus infection. From these results, *Temperature* is the only risk factor that differentiates between coinfection and single infections.

4. Predictive analysis

In this section we aim to propose a methodology that can help make timely decisions for targeted treatment in pathogens coinfection cases. We show that we can derive a predictive analysis to discriminate arbovirus positive and arbovirus negative cases among coinfected patients.

4.1. Testing independence between arbovirus and malaria

In the multinomial model given by (1) in Section 3.1, we can test the independence between arboviral and malaria infections.

The joint statistical distribution of arboviral infection (A^+) and malaria infection (M^+) is given in Table 4. As in Table 1 and Table 2, A^+ corresponds to an individual belonging to categories 1 or 3 of the response Y, and M^+

	$A=0$	$A=1$	law of M^+
$M=0$	π_0	π_1	$\mathrm{P}(M^+=0)=\pi_0+\pi_1$
$M=1$	π_2	π_3	$\mathrm{P}(M^+=1)=\pi_2+\pi_3$
law of A^+	$\mathrm{P}(A=0)=\pi_0+\pi_2$	$\mathrm{P}(A=1)=\pi_1+\pi_3$	1

Table 4: Joint distribution of arboviral infection and malaria infection

corresponds to an individual belonging to categories 2 or 3 of the response Y.
Independence between arboviral and malaria infections means that for all (l_1, l_2) in $\{0,1\}$,

$$\mathrm{P}(M^+=l_1, A^+=l_2)=\mathrm{P}(M^+=l_1)\times\mathrm{P}(A^+=l_2).$$

The independence hypothesis can be written in terms of parameters as:

$$H_0: \qquad \text{“}\beta_3=\beta_1+\beta_2''.$$

The Wald statistic to test H_0 against its two-sided alternative is computed as

$$W=h(\widehat{\beta})^T\Sigma^{-1}h(\widehat{\beta}),$$

with $h(\widehat{\beta})=\widehat{\beta}_3-\widehat{\beta}_1-\widehat{\beta}_2$ and $\Sigma=DVD^T$ where $D=(-Id_{p+1},-Id_{p+1},Id_{p+1})$; Id_p is the $p\times p$ identity matrix and V is an estimator of the variance of $\widehat{\beta}=(\widehat{\beta}_1,\widehat{\beta}_2,\widehat{\beta}_3)^T$. Under H_0, W is asymptotically distributed as a chi-square variable with $(p+1)$ degrees of freedom. Under H_1, W converges to infinity as the sample size goes to infinity.

We fitted model (1) including the eight covariates selected in Section 3.3 and we computed the independence test. Based on IgM/IgG data, the independence hypothesis was rejected with a p-value equal to $1.46.10^{-6}$. 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.
Applying the test on the IgM data, we computed the 1000 p-values corresponding to the 1000 sub-samples and obtained that 42.5% of them were less than 0.05. So we can not reject the independence hypothesis in a majority of sub-samples. It can be explained by the fact that the size of the sub-samples

is small (139) and the asymptotic approximation of the law of the test statistic is not accurate. Moreover, the IgM data set may not contain enough information to explain coinfection. Which means that the independence test lacks of power.
In the following, we will only consider the IgM/IgG data set to propose a predictive analysis.

4.2. Diagnosis of arboviral disease

In this section, we present a methodology to help doctors to diagnose the arboviral infected patients whose symptoms are masked by malaria symptoms. We propose to base the diagnosis on the conditional probability $\mathrm{P}(C|M)$ to be coinfected given that malaria infection is observed. This probability is the quantity of interest because arboviral infections are considered by healthcare workers only if malaria tests are negative. In absence of rapid arbovirus detection tests, the aim is to provide a decision support tool to determine if an arbovirus could be responsible for the clinical symptoms of the patient coinfection.
Based on the previous results of the IgM/IgG data set, the independence test of Section 4.1 displays an association between malaria and arboviral infections. Then the probability $\mathrm{P}(C|M)$ can be computed in function of the π_k probabilities estimated from the multinomial logit model. For an individual with covariate x,

$$\mathrm{P}(C|M) = \frac{\widehat{\pi}_3(x)}{\widehat{\pi}_3(x) + \widehat{\pi}_2(x)} = \frac{e^{\langle x, \widehat{\beta_3} \rangle}}{e^{\langle x, \widehat{\beta_3} \rangle} + e^{\langle x, \widehat{\beta_2} \rangle}}.$$

This probability can be used to differentiate whether the illness to be treated should be arbovirus or malaria. We propose a binary classification rule and we predict an arbovirus illness if the estimated coinfection probability is greater than a threshold value γ:

$$\begin{cases} \text{If } \mathrm{P}(C|M) \geq \gamma : & \textit{arbovirus positive case}, \\ \text{If } \mathrm{P}(C|M) < \gamma : & \textit{arbovirus negative case}. \end{cases}$$

The evaluation of the classification is based on the confusion matrix and the overall classification accuracy. The confusion matrix is used to compute true arbovirus positives (TP), false arbovirus positives (FP), true negatives

(TN) and false negatives (FN). A global performance measure is the miss-classification rate (MCR) defined as:

$$\text{MCR} = \frac{FP + FN}{N},$$

with $N = TP + FP + TN + FN$.
The analysis performed in this section is 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 9 that the optimal threshold is around $\gamma = 0.5$. Five-fold cross-validation was run different 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 5. 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

Predicted / ***True***	0	1
0	211	29
1	114	23

Table 5: Confusion table with $\gamma = 0.5$.

strategies can 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 mistakes 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 6 that the number of FN is reduced but the MCR remains too high (46.7%).

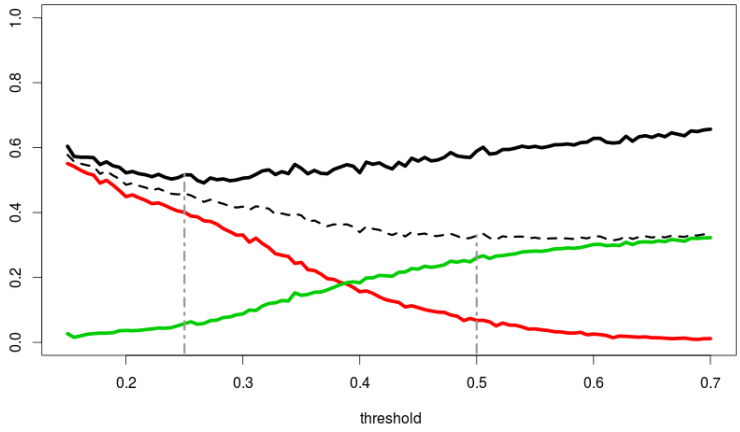

Figure 9: IgM/IgG data: estimated cross-validation miss-classification rate. The WMCR is shown in black as a full line. The MCR is shown as a black dotted line. Increasing γ increases the number of FN (green full line) and decreases the FP (red full line).

True \ ***Predicted***	0	1
0	88	152
1	24	113

Table 6: Confusion table with $\gamma = 0.25$.

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 3.3.1 that these two variables are mostly indicative of arboviral disease. The threshold values were again chosen to minimize the WMCR using cross-validation. Table 7 gives the corre-

sponding results: the MCR is decreased to 36% while the number of FN is smaller than the number of FN of Table 5 and the number of TP is doubled.

True \ ***Predicted***	0	1
0	190	50
1	85	52

Table 7: Confusion table with $\gamma = 0.25$, $Age = 10$ and Number of sick days $= 3$.

The objective of these predictions was to assign patients to either a "Malaria" group or a "Arbovirus" group and to handle mystifying cases due to the similarity of the initial symptoms in both diseases. The classification procedure is based on the computation of the conditional probability $\mathrm{P}(C|M)$. The threshold parameter γ is calibrated to minimize the weighted misclassification rate. To improve the accuracy of the classification, we propose to take advantage of the covariates that were selected in Section 3.3.1 as arbovirus specific covariates. Based on these specific covariates, we filtered the positive predicted patients and obtained better results.
The performance of a classification procedure is greatly affected by the quality of data source. We based our analysis on the IgM/IgG data collected from patients who were tested positive to IgM or IgG. As IgG antibodies appear later in time in blood than IgM antibodies, they may not reflect a current infection, thus minimizing the possibility of finding a true correlation with the current recorded symptoms. This drawback may reduce the prediction capacity of the classification procedure. Also, false positives and false negatives from biological tests could impact these results. However, the diagnostic tests used in the study displayed high sensitivity and specificity parameters; then their impact on the results should be negligible.

5. Discussion

Misdiagnosis of arbovirus coinfections as malaria infections may increase the spread of arbovirus diseases in areas where fast diagnostic assays are not available. This study proposes an appropriate statistical methodology that can assist doctors in the elaboration of the differential diagnosis of febrile cases for arboviruses.

Our analysis is based on a real-life medical data set. In the original IgM data set, arbovirus positive individuals are identified as individuals likely to be in the early stages of arbovirus illness. It is the relevant data set for the classification problem. However, the positive cases constitute only a very small minority class of the data (39 positive cases over 12288 individuals). Several sampling strategies have been developed to learn from imbalanced data sets [18] and to correct classification of the rare class. [16] proposed a categorisation of these approaches into two main categories: data pre-processing and modifications on the learning algorithms. Algorithm level strategies including random forest solutions have been discussed to deal with the imbalanced data classification problem ([19], [20]). They require a good insight into the modified learning algorithm and a precise identification of reasons for its failure in mining imbalanced data sets. As we were interested in the statistical methodology that could be applied to a more relevant data set, we took solutions that pre-process the given imbalanced data set. Since the data distribution is changed to make standard methods focus on the cases that are more relevant for our problem, the results should be interpreted carefully.

To analyze coinfection data we propose a methodology with three steps: 1. a variable selection with random forests; 2. an analysis of the influent factors through multinomial model fitting and odd ratios computation; 3. a predictive analysis based on coinfection probabilities. From our experiments, we can say that the random forests algorithm is a robust method to select the important variables for the different diseases. The analysis of the odd ratios allows to identify the risk factors that characterize each disease. We observed that higher values of number of sick days and of age are mostly indicative of arboviral disease while higher values of temperature and presence of nausea or vomiting symptoms during the rainy season are mostly indicative of malaria disease. The results also pointed out that a high-grade fever could be considered as a differential diagnostic for malaria and arbovirus coinfection, which is in agreement with the study of [13]. The classification rule based on coinfection probability, age and number of sick days identifies coinfected patients to be treated for arbovirus with global accuracy of 65%. The results could be improved on a more suitable data set. A future study will apply this methodology to coinfection data between malaria and other pathogens more easily detectable in the early stages of infection than arboviruses.

Acknowledgements : The authors thank two referees for detailed and

helpful comments that improved the manuscript.

References

[1] M. K. Mohapatra, P. Patra, R. K. Agrawala, Manifestation and outcome of concurrent malaria and dengue infection., Journal of Vector Borne Diseases 49 (4) (2012) 262–265.

[2] M. Mushtaq, M. Qadri, A. Rashid, Concurrent infection with dengue and malaria: An unusual presentation., Hindawi Publishing cooporation Case report in Medecine 2013 (2013) 2. doi:ArticleID520181,2.

[3] B. Carme, S. Matheus, G. Donutil, O. Raulin, M. Nacher, J. Morvan, Concurent dengue and malaria in cayenne hospital, french guiana, Emerging Infections Diseases 15 (4) (2009) 668–671. doi:10.3201/eid1504.080891.

[4] C. S. Arya, L. K. Mehta, N. Agarwal, K. A. Bharat, G. Mathai, A. Moondhara, Episodes of concurrent dengue and malaria, Dengue Bulletin 29.

[5] S. Deresinski, Concurrent plasmodium vivax malaria and dengue., Emerging Infectious Diseases 12 (11) (2006) 1802.

[6] N. Ali, A. Nadeem, M. Anwar, W. U. Z. Tariq, R. A. Chotani, Dengue fever in malaria endemic areas., Journal of theCollege of Physicians and Surgeons–Pakistan: JCPSP 16 (5) (2006) 340–342.

[7] N. Senn, D. Suarkia, D. Manong, P. M. Siba, W. J. H. Mcbride, Contribution of dengue fever to the burden of acute febrile illnesses in papua new guinea: An age-specific prospective study., The American Journal of Tropical Medecine and Hygiene 85 (1) (2011) 132–137. doi:10.4269/ajtmh.2011.10-0482.

[8] R. N. Charrel, P. Brouqui, C. Foucault, X. De Lamballerie, Concurrent dengue and malaria., Emerging Infections Diseases 11 (7) (2005) 1153–1154.

[9] V. R. Mendonça, B. B. Andrade, B. M. L. Souza, Ligia C L adn Magalhães, M. P. G. Mourão, M. V. G. Lacerda, M. Barral-Netto, Unravelling the patterns of host immune responses in plasmodium vivax malaria and dengue co-infection. 14:315. doi:10.1186/s12936-015-0835-8.

[10] M. Baba, C. H. Logue, B. Oderinde, H. Abdulmaleek, J. Williams, J. Lewis, T. R. Laws, R. Hewson, A. Marcello, P. D' Agaro, Evidence of arbovirus co-infection in suspected febrile malaria and typhoid patients in nigeria., The Journal of Infection in Developing Countries 7 (1) (2013) 51–59.

[11] S. Thiam, M. Thior, B. Faye, M. N diop, M. L. Diouf, M. B. Diouf, I. Diallo, F. B. Fall, J. L. Ndiaye, A. Albertini, E. Lee, P. Jorgensen, O. Gaye, D. Bell, Major reduction in anti-malarial drug consumption in senegal after nation-wide introduction of malaria rapid diagnostic tests, PloS One 6 (4) (2011) 1–7.

[12] ANSD, programme national de lutte contre le paludisme au Senegal. (2009).

[13] A. Sow, C. Loucoubar, D. Diallo, O. Faye, Y. Ndiaye, C. S. Senghor, A. T. Dia, O. Faye, S. C. Weaver, M. Diallo, D. Malvy, A. A. Sall, Concurrent malaria and arbovirus infections in kedougou, southeastern senegal, Malaria Journal 15:47. doi:10.1186/s12936-016-1100-5.

[14] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests., Pattern Recognition Letters 31 (14) (2010) 2225–2236.

[15] L. Breiman, Random forest, Machine Learning 45 (2001) 5–32.

[16] P. Branco, L. Torgo, R. P. Ribeiro, A survey of predictive modeling on imbalanced domains, ACM Computing Surveys (CSUR) 49 (31) (2016) 31:1–31:50. doi:10.1145/2907070.

[17] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Vsurf: An r package for variable selection using random forests., The R Journal 7 (2) (2015) 19–33.

[18] C. Chen, A. Liaw, L. Breiman, Using random forests to learn imbalanced data., Tech. rep., University of Berkeley (2006).

[19] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (4) (2016) 221–232. doi:10.1007/s13748-016-0094-0.

[20] N. V. Chawla, Data mining for imbalanced datasets: An overview., In: O. Maimon, L. Rokach (eds.) Data Mining and Knowledge Discovery Hand-book, SPRINGER (2010) 875–886.