Supplementary: The statistical model and inference

# The model

This is a model borrowed from econometrics (1) and recently used in the context of infectious diseases which occur simultaneously or are inherently influenced by correlated binary outcomes. We adapted the methodology in the framework of these two binary highly correlated outcomes which can be associated with a similar (or not) set of covariates and the presence of one is highly correlated with the other. A lay description of the approach is presented in the main text. Here, we give details on the mathematical equations and the statistical inference.

Let us denote by X and Y the two binary random variables indicating the presence/absence of an individual to DENV and CHIKV, respectively. This pair of responses define a binary bivariate outcome denoted by (X, Y).

Let us also denote

 vector of variables associated with

 vector of unknown parameters which quantify the effect of variables associated with

 vector of variables associated with

 vector of unknown parameters which quantify the effect of variables associated with ,

 , ,

NB: We deliberately keep the generality of the model specification although in our scenario

Then, the two simultaneous logistic equations which depict the nonlinear association with age and the rest of the other variables are written in terms of a system of conditional probabilities on the potential explanatory variables and their random effects:

(1)

The triple index ( – denotes individual in the family and the ward

 denote the age-group constants associated weights (see the subsequent section on Population adjusting weights)

 denote the ward constants associated weights (see the subsequent section on Population adjusting weights)

 denote the random effect associated with the ward

 denote the random effect associated with the family

The latter two sets of random effects are assumed to be drawn from a two normal bivariate normal distributions (MVN) which model the shared and unshared similarities and account for unmeasured aspects such as mosquitos’ burden around the households or weather patterns for instance.

On one hand side, from the first simultaneous equation specification (1):

= (2)

From the second simultaneous equation specification:

= (3)

Hence, using the key assumption that and keeping the fixed effects only for the sake of simplicity, the system can be written as

This implies that, if the occurrence of dengue makes more likely (or less likely) the occurrence of chikungunya, the converse is also necessarily true and is deemed as a nuisance parameter

This crucial assumption and elementary rules of conditioned probabilities leads to the following four statements derived from

Hence, combining either (4) + (7) (or (5) + (6)) by multiplication

But all four probabilities situations ((1, 0), (0, 1), (1, 1), (0,0)) sum to 1:

So

=1

Therefore, the predicted joint probabilities can be derived as it follows:

Denoting by

If the vector of the unknown parameters are denoted by **,** the model likelihood can be formally written as

Bayesian inference consist of summarizing the posterior distributions of the model parameters and requires the joint probability distribution over all random quantities, parameters or observables. The marginal and conditional probability distributions of the vector of unknown parameters are denoted by and . The posterior distribution is proportional to the likelihood multiplied by their prior distributions – all assumed non informative as stated in the text. That is

# Statistical Inference

The simultaneous equations systems (1) define a probability model which has been implemented in OpenBUGS - which makes this approach very appealing even for larger datasets (OpenBUGS program for the simplest model appended). A Bayesian framework for statistical inference using Markov chain Monte Carlo (MCMC) estimation methods naturally accommodates missing responses (5%) in the data and allows straightforward assessment of uncertainties in nonlinear models under missing at random assumption. Sensitivities analyses to the missing data (by assuming all positives and all negatives) have been conducted but the results were not altered.

The posterior conditional odds ratios (OR) distributions are presented in Table 2. However, for an easily digestible reading, the summary statistics are presented in terms of means and standard deviations of the predicted joint probabilities by age and stratified by different socio-economic factors that we accounted for.

The MCMC estimation followed 1 million iterations and the corresponding chains (log scale) are displayed in Supplementary Figure 8 – including that of the nuisance parameter. Supplementary Figure 9 display the correlation between the parameters which are all lower than 0.5 in absolute values with the exception of that between the constant and age corresponding to the dengue equations with appear to be highly negatively correlated (-0.87).

Non-informative priors have been used for parameter estimation.

# Population adjusting weights used in modelling

We used the 2010 Census data (the latest) adjusted for an average annual growth of in the Ecuador population. All statistical analyses should ultimately serve population predictions. The post-stratification weights, namely calibration, is a technique which corrects the estimation for population misrepresentation which often occurs in ad-hoc surveys or when researchers use already recruited participants. One set of weights are aimed at reducing the effect of Rosa Zarate where most of the data were collected and at inflating the influence of the misrepresented groups (La Union and 10-19 years) on the model likelihood. These weights are associated with each individual and are additive for the log scale model formulation of the model (see equations above). The other set of weights are aimed at reducing the effect of those in young group and inflate the effects of older people less represented in the sample. The weights are displayed Supplementary Tables 3 and 4. Each observation is associated with a weight according to individual’s age group and one by the household location and their additive effect is accounted for in the predictors (log scale so their effect on the actual odds ratios or odds would be multiplicative).

# Goodness of fit

The model is developed using a fully parametric paradigm and inherits complexities including random effects for binary outcomes which are not associated with any classical or formal goodness of fit measures.

However, in reference to the basic independent binary outcomes, the only goodness-of-fit measure following a potential *individual based model* fit to these data is that of Hosmer-Lemeshow statistic. That invokes partition of the data *into groups* (<https://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/>). That is because the individual level raw observations are either 0 or 1 whilst the individual level predicted values can be anywhere between 0 and 1 and hence a one-to-one observed-predicted plot often used for continuous data loses its meaning.

Hence, our approach alludes to that of basic logistic regression and follows similar lines to those

developed in Chis Ster, I. Journal of Applied Statistics, 2012. **39**(11): p. 2523-2542, (*Inference*

*for serological surveys investigating past exposures to infections resulting in long-lasting*

*immunity – an approach using finite mixture models with concomitant information)* within a

related and yet equally complex modelling framework.

We split the data into groups, by 5 years age-groups and geographical sub-regions, and calculated the average predicted probabilities and their 95%CrIs for the outcomes (((D+, C-), (D-, C+), (D+, C+) and also for all dengue and chikungunya) within each group. Then we calculated the expected number of positives corresponding each group and each outcome component by multiplying the corresponding average predicted probabilities with the total number of people within that group. We plotted the observed numbers vs. the predicted numbers in each group.

Given the full parametric, parsimonious nature of this model, and the average sized data – we deemed our model as a good fit to these data.

# OpenBUGS code:

model{

 #model's likelihood

for (i in 1:319){

 log\_age[i]<-log(age\_final\_years[i]);

 log\_dist[i]<-log(1+min\_dist[i]/250);

 weights\_final[i]<-weights[i]+weight\_age[i];

 dengue[i]~dbern(p[i,1]);

 chick[i] ~dbern(p[i,2]);

 eta[i,1]<-beta1[1]+beta1[2]\*log\_age[i]+alpha\*chick[i]+beta1[3]\*gender[i]+

 beta1[4]\*ses\_bin[i]+

 weights\_final[i]+

 theta[family[i],1]+

 ward\_level[ward\_index[i],1];

 eta[i,2]<- beta2[1]+beta2[2]\*log\_age[i]+alpha\*dengue[i]+

 beta2[3]\*gender[i]+

 beta2[4]\*urbanrural[i]+beta2[5]\*log\_dist[i]+

 weights\_final[i]+

 theta[family[i],2]+

 ward\_level[ward\_index[i],2]; logit(p[i,1])<-eta[i,1];

 logit(p[i,2])<-eta[i,2];

 }

 for (j in 1:151){ theta[j,1:2] ~ dmnorm(mu.fam[],R\_fam[,]);

 }

 for(j in 1:49){

 ward\_level[j,1:2] ~ dmnorm(mu.ward[],R\_ward[,]);

 }

 for (k in 1:2){

 mu.fam[k] <-0.0;

 mu.ward[k] <-0.0;

 }

 R\_fam[1:2,1:2] ~dwish(Omega\_fam[,],2);

 R\_ward[1:2,1:2]~dwish(Omega\_ward[,],2);

 Omega\_fam[1,1]<-1; Omega\_fam[1,2]<-0; Omega\_fam[2,1]<-0; Omega\_fam[2,2]<-1;

 Omega\_ward[1,1]<-1; Omega\_ward[1,2]<-0; Omega\_ward[2,1]<-0; Omega\_ward[2,2]<-1;

 Sigma.fam[1:2, 1:2] <-inverse(R\_fam[,]);

 Sigma.ward[1:2, 1:2]<-inverse(R\_ward[,]);

 for (j in 1:2){

 sigma.fam[j]<-sqrt(Sigma.fam[j,j]);

 sigma.ward[j]<-sqrt(Sigma.ward[j,j]);

 for(k in 1:j){

 rho\_fam[j,k]<-0;

 rho\_ward[j,k]<-0;

 }

 for(k in (j+1):2){

 rho\_fam[j,k]<-Sigma.fam[j,k]/(sigma.fam[j]\*sigma.fam[k]);

 rho\_ward[j,k]<-Sigma.ward[j,k]/(sigma.ward[j]\*sigma.ward[k]);

 }

 }

 alpha ~ dnorm( 0.0, 0.00001);

 for (j in 1:4){

 beta1[j] ~ dnorm( 0.0, 0.00001);

 beta2[j] ~ dnorm( 0.0, 0.00001);

 exp\_beta1[j]<-exp(beta1[j]);

 exp\_beta2[j]<-exp(beta2[j]);

 eff1[j] <-step(beta1[j]);

 eff2[j] <-step(beta2[j]);

 }

 exp\_alpha<-exp(alpha);

 beta2[5] ~ dnorm( 0.0, 0.00001);

 exp\_beta2[5]<-exp(beta2[5]);

}

# References

1. Schmidt P, Strauss RP. Estimation of Models with Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach. Econometrica. 1975;43(4):745-55.
2. Ghebremichael M. Joint modeling of correlated binary outcomes: HIV-1 and HSV-2 co-infection. Journal of Applied Statistics. 2015 2015/10/03;42(10):2180-91.
3. Goldstein H. Multilevel Statistical Models. 4th Edition ed: JohnWiley and Sons, Ltd, Chichester; 2011.