IJE vol.33 no.6 © International Epidemiological Association 2004; all rights reserved.
Advance Access publication 8 July 2004

International Journal of Epidemiology 2004;**33**:1337–1344
doi:10.1093/ije/dyh274

Multilevel modelling and malaria: a new method for an old disease

F Mauny,[1,3] JF Viel,[1] P Handschumacher[2] and B Sellin[3]

Accepted 2 June 2004

Background Malaria is influenced by a web of individual and ecological factors, i.e. factors relating to people and relating to environment. For a long time analysing these factors concurrently has raised statistical problems. Multilevel modelling provides a new attractive solution, which is still uncommon in tropical medicine.

Methods Using an actual data set of 3864 individuals from 38 villages of the Highland Madagascar, a two-level modelling process is presented. Individual malaria parasitaemia is modelled step by step according to age (individual factor), altitude, and DDT indoor house-spraying status (village factors).

Results The hierarchical organization of a data set in levels, fixed and random effects, and cross-level interactions are considered. Accurate estimations of standard errors, impact of unknown or unmeasured variables quantified and accounted for through random effects, are the highlighted advantages of multilevel modelling.

Conclusion While not denying the importance of understanding an aetiological chain, the authors recommend an increased use of multilevel modelling, mainly to identify accurately ecological targets for public health policy.

Keywords Multilevel model, malaria, individual variable, ecological variable, Madagascar

The environment represents the third dimension of the epidemiological triad: person, time, and space. Depending on the scale, the environment can be defined at different levels and characterized by specific factors. A house may be characterized by the type of the roof, the number of sleeping rooms, whether or not animals are sleeping in the house. A village may be described by the proximity of irrigated lands or the presence of a village health educator. Districts may or may not be involved in a bed net programme. Although well identified, quantifying the relative influence of each of these environmental factors (and others) in malaria transmission would raise serious methodological difficulties.

In medical research, the environmental dimension has been neglected in favour of an individual-centred approach.[1–3] In recent years, the question of whether and how environmental factors could have significance for health has been increasingly explored.[4–9] The influence of the social environment on individual health outcome was for example reported for low birthweight,[10] diastolic blood pressure, or all-cause mortality.[11]

Until recently, it was necessary to choose between the individual-centred and the collective-centred (also called ecological) approach for methodological reasons. In the collective approach, therefore, spatial analytical methods and geographical information systems explore diseases at a supra-individual aggregated level.[12] Numbers of cases, and prevalence or incidence rates are related to geographical units, and ecological exposure estimations for comparative or predictive purposes are composed on the same scale. In parasitological field research, these methods are increasingly used, notably for malaria.[13–16]

Although useful, spatial analytical methods could reduce the scope of an investigation since exposure and characteristics of each of the individuals are not taken into account. Indeed, the origin of variation between areas could be explained by a complex combination of factors which are characterizing people (the individual level) or areas (the group level). When an individual factor is a characteristic of subjects who are more likely to be ill, variability of its distribution across areas will influence health outcomes in a given area: this is called a composition effect. So, relations between individual and 'supra-individual' determinants are of particular interest, especially for investigating the reasons of variation between

[1] Department of Public Health, Biostatistics and Epidemiology Unit, Faculty of Medicine, 2, place Saint Jacques, 25030 Besançon Cedex, France.

[2] Institut de Recherche pour le Développement, Institut de Géographie, 3, rue de l'Argonne, 67083 Strasbourg Cedex, France.

[3] Programme RAMSE, Institut de Recherche pour le Développement, BP 434, Antananarivo, Madagascar.

Correspondence: Dr Frederic Mauny, Department of Public Health, Biostatistics and Epidemiology Unit, Faculty of Medicine, 2, place Saint Jacques, 25030 Besançon Cedex, France. E-mail: frederic.mauny@ufc-chu.univ-fcomte.fr

areas: are the people living in the areas different or are the areas different, i.e. is it a composition or a context effect?[17]

Connecting individual and collective exposures necessitates analysis of several collective situations simultaneously, and in each one, several individuals. Gathered in the same situation (household, village …), individuals are more similar to each other than individuals from different contexts. They are organized into groups of dependent data (also called clusters); individuals are said to be nested within household, village, or geographical areas. Nesting is also known as 'design effect' or 'data structuring'. Consider a hypothetical study to measure enteric helminth infection in 200 people in each of two villages, one with piped water and one with water coming from the local stream. Without accounting for nesting statistics assume $n = 400$, whereas for the comparison between piped water versus non-piped water $n = 2$. Such a data set raises, in statistical terms, the issue of correlated data analysis.[18] This dependence between observations is contrary to one of the basic assumptions of the conventional regression technique, i.e. independence of observations. Assuming a size of independent observations that is inappropriate, the statistical analysis will be wrong. A recent illustration from biological literature was provided by Morisson.[19]

For a decade, a new statistical approach based on multilevel modelling has been available, aided by the increase in computing power. A variety of names have been used synonymously for 'multilevel model': 'hierarchical model', 'random effect model', 'variance component model', or 'mixed model'.[20] First widespread in social sciences, many multilevel modelling studies are now published in health sciences[6,21–23] but it is noteworthy that few deal with infectious or parasitological diseases.[24,25] The principle of multilevel modelling is to analyse simultaneously the influence of individual factors and environmental factors. The data set is structured as a succession of nested levels: people are gathered by house, houses are gathered by village, villages are gathered by district … Outcomes defined at the lowest level (parasite burden of each people) are then modelled as a function of variables characterizing the different levels (people, house, village, district).

The aim of this paper is to demonstrate multilevel modelling in malaria and to show how misleading an analysis can be if it considers only one level. To this end, an actual malaria data set is used to illustrate the main outlines of such an approach.

Population and Methods

Malaria, as many other parasitological diseases, really embodies diseases influenced by a web of determinants defined at different levels. Vector breeding, exposure to transmission, immunity, morbidity, clinical expression, drug resistance, prevention, are all subject to a wide variability. Most of them are modulated by both individual and environmental determinants: individual response to a collective exposition, migration, compliance to health programmes.

Study area

In Central Highland Madagascar transmission of malaria is seasonal with morbidity decreasing from warm (January–May) to cold season (July–August).[26] From East to West, altitude declines, transmission period becomes longer, and malaria tends to be stable.[26,27] Within the area, more than 90% of the malaria infections are *Plasmodium falciparum*.[28] After the deadly epidemics of 1986–1988, in 1993 the Malagasy government started a 5-year indoor DDT house-spraying programme in areas located at altitudes between 1000 and 1500 m. Actually, 2 years after the campaign began, DDT was not sprayed in the all planned villages, and conversely DDT was sprayed in few villages which were not targeted by the programme. The main reasons for this situation were inaccessibility during the spraying period, missing product, and altitude misclassification.

Data set

It consists of a sub-sample extracted from a wide cross-sectional community-based study conducted in the Middle West of Madagascar in July 1995 by the RAMSE programme, a research programme involving the Malagasy Health Ministry, the French Institut Pasteur, and the French Institut de Recherche pour le Développement (formerly ORSTOM). Individual and collective data were collected according to standardized field procedures for questionnaires, clinical examinations, and biological sampling. Informed consent was obtained from all adult participants and from parents or legal guardians of minors. Inhabitants of the 38 villages located in the area covered by the DDT spraying programme were selected for analysis. After exclusion of 238 individuals (infants or missing values), 3864 subjects comprised the final data set.

Outcome and factors

The individual health outcome we considered is the presence or absence of *Plasmodium* in blood samples. Aggregated results are expressed as parasite prevalence, i.e. the percentage of *Plasmodium* positive subjects. The factors used to explain outcomes are defined at two levels: individual (age) and village (DDT-spaying status and altitude, range 900–1600 m). Variables were coded as follows. Age and altitude were split into two categories (thresholds of 10 years and 1300 m, respectively). DDT-spraying was coded as unsprayed or sprayed once or more since the beginning of the campaign.

Multilevel modelling

The outcome (carriage of *Plasmodium*) is a binary variable (*Plasmodium*: yes/no). Logistic models were used to assess the influence of independent variables on the odds of being *Plasmodium* positive. Let π_i be the predicted probability (and $\frac{\pi_i}{1-\pi_i}$ the odds) of being *Plasmodium* positive for the *i*th individual, the logit function is defined as follows:

$$\operatorname{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

and the equation of a conventional logistic model is:

$$\operatorname{logit}(\pi_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where β_0 is the intercept, and $\beta_1 \dots \beta_p$ are the regression coefficients of independent variables $X_1 \dots X_p$. The odds ratio (OR) associated with the variable X_1 is the exponential function of its parameter β_1 ($OR_{X1} = exp(\beta_1)$).

In the current analysis, a multilevel statistical approach was used to model the relation between malaria and three independent factors. Two levels of organization were stated (individual and village) in a multilevel logistic regression model. Let π_{iv} be the predicted probability of being *Plasmodium* positive for the *i*th individual of the *v*th village. The logit function becomes:

$$\text{logit}(\pi_{iv}) = \log\left(\frac{\pi_{iv}}{1 - \pi_{iv}}\right)$$

The general equation of a multilevel logistic model is:[29]

$$\text{logit}(\pi_{iv}) = \beta_{0v} + \beta_1 X_{1iv} + \cdots + \beta_p X_{pv}$$

The key difference between conventional (single level) and multilevel models is the structure of the random part of the model which is also called residual variation or error. In the conventional model, there is only one level and the structure of the residual variation is reduced to one value: the individual-level residual variance. In the multilevel model, the structure of the random part (residual variation) is more complex and partitioned among levels of the data hierarchy. Here, the random part of the logistic model is partitioned among an individual level variance (which is set to be Binomial) and village level variance.

From a computational point of view, multilevel modelling can be seen as a two-stage process.[20] First, a separate individual-level regression is defined for each village. Then, each of the village-specific coefficients are modelled as a function of village variables. So, multilevel analysis allows the partition of the village-specific coefficients: a fixed part that is common across villages and a random part varying between villages. Coefficients in the models were estimated using a Second Order Penalised Quasi Likelihood (PQL).[29,30] Fixed and random coefficients were successively estimated, and iterative estimations were performed until the procedure converged. For non-Normal models, the likelihood statistic can only be approximated, so statistical significance of fixed parameters was tested using Wald 95% CI.[30,31] Normal distribution of the village-level residuals was graphically checked. The SAS package was used for conventional logistic modelling and the MlwiN software was used for multilevel modelling.[32]

Description of Levels

The overall parasite prevalence was 15%, modified by age (<10 years = 23%, ⩾10 years = 11%), altitude (<1300 m = 22%, ⩾1300 m = 6%), and DDT status (unsprayed = 31%, sprayed = 8%).

Cross-tabulation of the three determinants indicates a more subtle pattern (Table 1). Individuals are classified according to their age, the altitude, and the DDT-spraying of their village. Overall parasite prevalence represents the average prevalence of the sub-group in question. It reflects an approach focusing on the individual, and considers the subjects statistically independent from each other. If we expect that environment could modify the probability of being *Plasmodium* positive, then this approach focusing on the individual implies that each of the 3764 people lives in an environment independent from the environment of the other subjects. Conversely, village parasite prevalence represents a collective approach (here focusing on village). Here, the environment is homogeneous at the scale of the villages. When comparing prevalences expressed by sub-group and by village, the relationship between the different factors appears to be complex. The two approaches seem to bring complementary information: trends, and deviations to trends.

In Figure 1 the age-specific parasite prevalence is plotted for each of the 38 villages. Villages are ranked by ascending altitude (x-axis). DDT-spraying is indicated by the colour of the bars (unsprayed = white and sprayed = striped). The parasite prevalence is displayed on the y-axis. For a village, two age groups (1–9 years, ⩾10 years) are displayed on the z-axis from back to front, respectively. This Figure illustrates the great variability exhibited by these data. Main trends already suspected for altitude, DDT-spraying, and age are noticeable again. However, deviations from these trends are also pointed out. In other words, the three factors only explain a part of the variability between the bars. For a given altitude and DDT-spraying, between-village differences remain. Lastly, relative to the oldest group, parasite prevalence in the first age group

Table 1 *Plasmodium* prevalence[a] according to subjects' age classes, village altitude, and DDT-spraying

	Altitude				
	<1300 m		⩾1300 m		
	Unsprayed	Sprayed	Unsprayed	Sprayed	**All areas**
No. of villages	8	8	4	18	38
Age < 10 years					
No. of subjects	303	314	63	523	1203
Overall parasite prevalence	59.7	13.0	20.6	7.5	22.8
Range of village prevalence	43.7,82.4	0.0,17.4	0.0,40.7	0.0,54.6	0.0, 82.4
Age ⩾ 10 years					
No. of subjects	726	659	137	1039	2561
Overall parasite prevalence	23.4	8.5	12.4	4.1	11.2
Range of village prevalence	10.7,32.0	0.0,14.5	1.7,25.0	0.0,19.2	0.0, 32.0
All ages					
No. of subjects	1029	973	200	1562	3764
Overall parasite prevalence	34.1	10.0	8.8	5.1	14.9
Range of village prevalence	22.3,43.4	0.0,14.7	1.2,30.6	0.0,29.7	0.0, 43.4

[a] Prevalences are expressed as percentage of subjects *Plasmodium* positive.

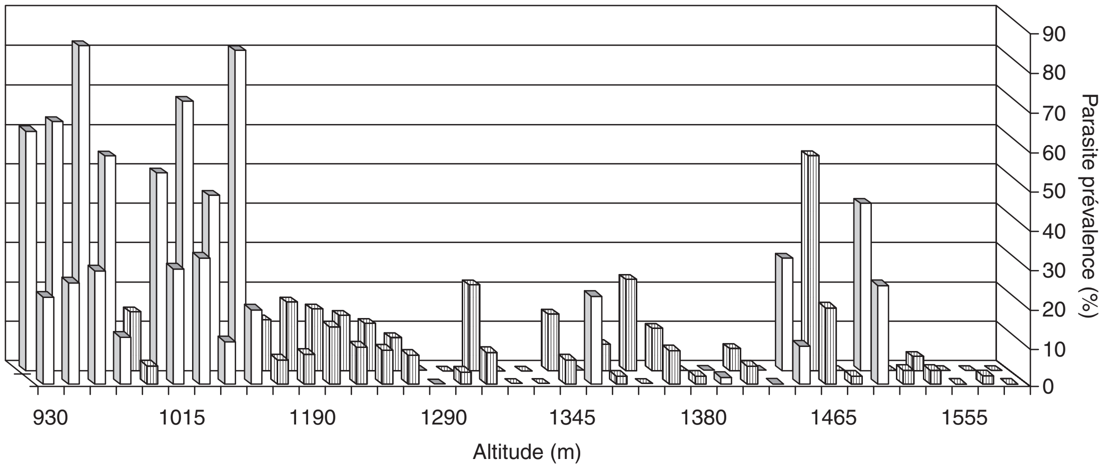

Figure 1 Parasite prevalence by age group (1–9, ⩾10) amongst the 38 villages according to altitude and DDT sprayed/unsprayed status. The colour of the bars indicates whether the village was sprayed (stripped) or unsprayed (white). For each village, two age classes are displayed on the z-axis (1–9 years = back box, ⩾10 years = front box)

(1–9 years) appears not to be uniform across villages, suggesting that influence of age (if any) may not be constant across villages.

Modelling Process

All the modelling parameters shown in Tables 2 to 4 are statistically different from 0.

Basic variance multilevel modelling

The equation of the null model—no variable introduced—is (Model A in Table 2):

$$\text{logit}(\pi_{iv}) = \beta_{0v} = \beta_0 + u_{0v} \qquad \text{(A)}$$

where β_0 is the 'average intercept', identical for the 38 villages. Thus, the model allows for residual variations about this intercept. Here, residual variations quantify differences between what is measured on average in the area and what is measured locally in each village. These differences, called village-level residuals and noted u_{0v}, are attributable to differences across village situations. They are assumed to be normally distributed, with mean zero. Their variance, σ^2_{0v}, represents the village-level variance. σ^2_{0v} estimation is 1.922 in Model A (noted Ω_A). This variance, statistically different from zero, reflects a between-village heterogeneity, regarding *Plasmodium* prevalence.

In a second stage, age is introduced (Model B), and the equation becomes:

$$\text{logit}(\pi_{iv}) = \beta_0 + \beta_1 age_{iv} + u_{0v} \qquad \text{(B)}$$

with $\beta_0 = -1.817$, $\beta_1 = -1.136$, and variance$(u_{0v}) = \Omega_B = 2.077$. As expressed by its odds ratio (OR = exp(−1.136) = 0.32), age greater than 10 years is associated with decreased odds of being *Plasmodium* positive.

In Models C and D, altitude and DDT-spraying are successively added. Model equations are the following:

$$\text{logit}(\pi_{iv}) = \beta_0 + \beta_1 age_{iv} + \beta_2 altitude_v + u_{0v} \qquad \text{(C)}$$

$$\text{logit}(\pi_{iv}) = \beta_0 + \beta_1 age_{iv} + \beta_2 altitude_v + \beta_3 ddt\_status_v + u_{0v} \qquad \text{(D)}$$

Table 2. Basic multilevel logistic models with successive introduction of explanatory variables

	Models **Coefficients (standard error)**			
Parameter	A	B	C	D
Fixed part:				
Intercept	−2.486	−1.817	−0.810	−0.018
Individual factor				
Age (>10 years)	—	−1.136 (0.109)	−1.140 (0.111)	−1.143 (0.110)
Village factors				
Altitude (>1300 m)	—		−1.788 (0.421)	−1.130 (0.329)
DDT-spraying (yes)	—	—	—	−1.601 (0.338)
Random part				
$\sigma2_{0v}$ (village-level variance)	1.922 (0.516)	2.077 (0.563)	1.298 (0.370)	0.626 (0.198)

Again, from the respective altitude and DDT-spraying parameter values in Model D, odds ratios can be calculated (0.32 and 0.20, respectively). Adjusted for age and for each other, altitude >1300 m and DDT-sprayed status are also identified as independently associated with lower odds of being *Plasmodium* positive.

Considering the three Models A, C, and D, the village-level variance decreases as village-level factors are introduced (as indicated in Table 2: $\Omega_A = 1.922$, $\Omega_C = 1.298$, and $\Omega_D = 0.626$, respectively). So, when accounting for altitude and DDT-spraying, the part of the variability which is relevant at the village level becomes lower. In other words, the village-level variance quantifies the part of the variability which is relevant at this level but not explained by village-level determinants already introduced in the model.[29]

Finally, the percentage of village-level variance explained by altitude (Model C) and by both altitude and DDT status

Table 3 Complex variance multilevel logistic model and conventional logistic model

	E-multilevel model			**F-conventional model**		
Parameter	Estimates (SE[a])	OR[b]	95% CI	Estimates (SE)	OR	95%CI
Fixed part:						
Intercept	−0.647			0.036		
Individual factor						
Age(>10 years)	−0.798 (0.189)	0.45	0.26, 0.65	−1.071 (0.102)	0.34	0.28, 0.42
Village factors						
Altitude (>1300 m)	−0.883 (0.309)	0.41	0.23, 0.75	−0.883 (0.124)	0.41	0.32, 0.53
DDT-spraying (yes)	−1.151 (0.308)	0.32	0.17, 0.58	−1.543 (0.110)	0.21	0.17, 0.26
Random part: (*village level variance*)						
σ^2_{0v} (Intercept)	1.487 (0.463)	-		-	-	
σ_{01} (covariance)	−0.840 (0.323)					
σ^2_{1v} (Age at random)	0.619 (0.269)	-		-	-	

[a] Standard error.

[b] Odds ratio.

Table 4. Multilevel logistic model including cross-level interactions

	G-Multilevel model	
Parameter	Estimates	SE[a]
Fixed:		
Intercept	0.301	
Individual factor		
Age(>10years)	−1.643	0.238
Village factors		
Altitude (>1300 m)	−1.580	0.485
DDT status	−2.178	0.494
Interactions		
Age*Altitude	0.660	0.350
Age*DDT status	1.001	0.332
Random part *(village-level variance)*		
σ^2_{0v} (Intercept)	1.325	0.420
σ_{01}(covariance)	−0.575	0.247
σ^2_{1v} (Age at random)	0.312	0.174

[a] Standard error.

(Model D) can be calculated as follow:

percentage of village-level variance explained by altitude

$$= [(\Omega_A - \Omega_C)/\Omega_A] \times 100 = [(1.922 - 1.298)/1.922] \times 100 = 32\%$$

percentage of village-level variance explained by both altitude and DDT-spraying

$$= [(\Omega_A - \Omega_D)/\Omega_A] \times 100 = [(1.922 - 0.626)/1.922] \times 100 = 67\%$$

So, in the Model D, 33% of the village-level variance remain unexplained, indicating that some unmeasured or unknown village characteristics could be missing.[29]

Complex variance multilevel modelling

In models B to D, it is assumed that the odds ratio for age does not vary across villages. We may relax this assumption by adding the age coefficient as a random variable at the village level (Model E, Table 3). A test to zero on this random parameter allows to test for null hypothesis that the odds ratio for age does not vary across villages.

$$\text{logit}(\pi_{iv}) = \beta_0 + \beta_{1v}\textit{age}_{iv} + \beta_2\textit{altitude}_v + \beta_3\textit{ddt\_status}_v + u_{0v} \text{ (E)} \quad \text{with} \quad \beta_{1v} = \beta_1 + u_{1v}$$

Here, a random coefficient for age (0.619) means that the coefficient β for age significantly varies across villages. The difference between the 'average' fixed relationships and the relationships in each village is noted u_{1v}. The mean of the u_{1v} is zero and the variance is equal to 0.619 in Model E.

Fixed and random parameters in multilevel modelling

Intercept (β_0) and coefficients associated with age (β_1), altitude (β_2), and DDT status (β_3) are the fixed part of the model. This part is used to estimate the strength of associations between individual plasmodial status and exposures. This strength is identical—fixed—over all the population. Conversely, the village-level variance defines the village-level random part of the model—the between-village variability not explained by fixed effects. This village-level random part invalidates the assumption of independence between individuals, and confirms the actual organization of the data set in more than a single level. The fixed effects represent the 'study area average' effects whereas the random part variance provides an estimate of what could be explained by each level.[17]

Comparison between multilevel and conventional modelling

Conventional logistic regression is performed on a single level of organization (individuals). Neither the village level, nor the correlated structure of the data is considered. Consequently, variability of coefficients across villages is not allowed by the modelling process, i.e. the random part defined at the village level does not exist. Table 3 shows the results obtained by the conventional logistic modelling (Model F). Coefficient values (and odds ratio) are relatively close to those estimated by multilevel modelling. The main difference lies in smaller standard errors in the conventional logistic regression. Considering all observations to be independent, conventional modelling assumes more information in the data than there actually is.[18] Consequently, standard errors based on an independence assumption are underestimated and 95% CI are too narrow when observations are, in fact, correlated. The risk is then to reject too often the null hypothesis, and so to conclude statistical significance too often.[29,33] A new variable (the population size of the villages) was introduced in models E and F, and analysis were conducted with both modelling processes. With the conventional model, the odds ratio is 1.21; the 95% CI (1.10, 1.35) does not include the value of one, and the variable appears statistically 'significant'. With the multilevel model, the odds ratio is 1.27 (95% CI: 0.96, 1.65), and the variable becomes 'non-significant'. In other words, modelling data without taking into account correlation between subjects seems to give more precision to estimations, but conclusions are based on a false underlying hypothesis.

Cross-level interactions

Two independent factors interact if the effect of one of the factors differs depending on the other. One can imagine that an individual-based factor effect can vary with a village-based characteristic. Two so-called 'cross-level interactions' were significant in our example: age*altitude and age*DDT-spraying. The final parameters are shown in Model G (Table 4). These results can be interpreted as follows: the influence of age appears stronger <1300 m of altitude ($\text{OR} = \exp(-1.643) = 0.19$) than above ($\text{OR} = \exp(-1.643 + 0.660) = 0.37$); the influence of DDT is greater for subjects < 10 years old than for the older subjects ($\text{OR} = \exp(-2.178) = 0.11$) and $\text{OR} = \exp(-2.178 + 1.001) = 0.31$, respectively). Interactions are commonly tested with conventional models, but cross-levels interactions can only be correctly analysed with multilevel modelling.[17]

Discussion

This paper attempts to provide support for the use of a sound statistical approach based on multilevel modelling. Although these models are more complex in theory and practice, and their application requires a good definition of the real hierarchical structure of the data, they permit combination of exposure to group and individual factors. This is crucial in infectious diseases. Indeed, individual risks depend not only on the status of the subjects but also on the status of the community in which they live, as illustrated by the protective effect of a vaccine also depending on the cover rate in the population. Sometimes, only group determinants demonstrate association with infection.[3,34,35]

Collective (here village) factors

A general difficulty for supra-individual determinants lies in the definition of the space the attention is focused on. This could be neighbourhood, communities, areas, but generally refers to a person's immediate residential environment. Most important is to choose a scale yielding geographical areas which characteristics may be relevant to the specific health outcome studied.[9] In this study, supra-individual characteristics were defined at the level of the village of residence. So, relative to subject's activity and mobility, this scale could hide a part of the real subject's environment. For example, subjects living at high altitude could be regularly infected during seasonal migration to lower altitude regions, smoothing the measured differences associated with altitude.

Interpreting results concerning area factors is complex because many dimensions and determinants may be interrelated.[3,9] Here, differences between villages could be due to bio-ecological or human factors. Indeed, temperature and rainfall, known as limiting factors for the malaria cycle and vector development in the Highlands, were not introduced in the models.[26] However, those factors are strongly correlated with altitude, which synthesizes climatic and vegetal conditions. As altitude decreases, environment becomes more favourable to malaria development. So, part of the altitude influence certainly reflects the influence of climatic factors on malaria. Similarly, human factors could be variations in population density, building or housing type, and behaviour, some of which known to be risk factors.[26,36] DDT-sprayed status is associated with a decrease in the individual odds of malaria. Furthermore, the protective influence of this collective factor had been shown to be modulated by an individual factor (the age). So, multilevel modelling allows assessment of health programmes, both at the collective and the individual level. The influence of individual factors on the effectiveness of collective programmes (limiting or promoting) can be precisely investigated.

Individual factors

The moderate protective influence of age on parasitaemia is in agreement with a low level of acquired immunity, already known in this population of Central Highland Madagascar.[26,27,37] Of course, other individual factors such as socio-economic or nutritional status, treatment, and associated or past diseases could be incorporated for better explanation of complex relations, but this would be beyond the scope of this paper.

Random effect, modification of effect

Village-level variance can be interpreted as heterogeneity across villages for the probability of being *Plasmodium* positive. This variability is not attributable to variables already introduced in the model. The random effect can then be considered as a composite surrogate of unknown or unmeasured supra-individual factors (village characteristics) influencing the variable of interest (parasitaemia).[29]

Focal interactions between man and environment are responsible for local variability. One main advantage of multilevel modelling is to pick up interactions; most noticeable in this context is the effect of age which differ according to altitude. So, modification of effect across groups could be modelled by both random effect (effect varying randomly) and cross-level interactions (modifications linked to fixed characteristics). These complex relations have to be explored, tested, and retained with regard to their significance. As a matter of fact, exploring the fact that individual characteristics do not play the same role from one group to another opens important possibilities of improving our understanding of individual and group behaviours, spontaneously and in response to risks.

Alternatives to multilevel approach in regression for correlated data

Another statistical approach had been developed to handle correlated data, without explicitly accounting for heterogeneity across groups. All the terms 'population-averaged models', 'marginal models', or 'covariance pattern models' refer to this approach.[18,20,38] The Generalized Estimating Equation is one method to fit this kind of models.[33] In contrast to multilevel models, these models do not provide direct estimates of the variance structure, but treat these as nuisance parameters.[38,39] Between-group variation, influence of individual-level or group-level factors on this variation, and sources of intra-group correlation are not examined.[20] Another interest lies in the fact that multilevel modelling does not need equilibrated data, particularly when analysing repeated measures from individuals.[40]

Various multilevel models

Only individual and village levels have been considered while household level could have represented an interesting intermediate level. For the sake of clarity, we preferred to fit only two-level models, although multilevel modelling permits the definition of several nested levels.[32] Moreover, cross-classified models could even be used. For example, a structure where children could be classified by village (the environment where they live) and by school (the environment where they study), giving a cross-classified structure, instead of a nested one. Last, as with conventional models, the relation between outcome and determinants could be analysed using different underlying distributions: Gaussian, logistic, Poisson, negative binomial, etc.

Various software packages

The two main packages specialized in multilevel modelling are MlwiN and HLM. Conversely to HLM, MlwiN has general facilities which can be accessed through drop-down menus, including a user interface designed for fully interactive use and integrated functions for data manipulation. To set up a model, an 'equation window' is used in which the user specifies the model in the format it is usually written. Major software packages (SAS, STATA, S-PLUS, SYSTAT) also provide procedures for fitting multilevel models.[41,42] Finally, MIXOR, MIXREG, and MLA are programmes available for free. Furthers detailed reviews of multilevel software packages can be found in ref. 43.

Process and insight summaries

Global variability (random part of the model) has been partitioned in an individual-level and a village-level variability. After controlling for an individual factor, village-level variability still remained which rules out a purely composition effect of this factor (model B). This village-level variability was then partially explained and reduced when taking into account village factors (model D). Furthermore, results show that the between-village variability could be partially explained by differences in the age influence across villages (model E). Finally, it has been stated that variables which were defined at different scale interacted with each others (model G).

Characteristics from different levels of the social organization were analysed simultaneously. Consequently, questions about the appropriate level of analysis are redundant. Valid estimates were produced by taking into account dependence between observations. Multilevel modelling highlighted the contextual richness and complexity which were suggested by the Figure. In particular the relative susceptibility of a social group (the youths) appeared to be modified by the context in which this group was living. A part of this context was explicit (altitude, DDT) and another part remains unmeasured or/and unknown, which could potentially open news hypotheses. To conclude, the key point is that multilevel modelling allows a demonstration of the independent effect of area/group characteristics from individual factors, and *vice versa*. While not denying the importance of understanding the aetiological chain, identification of environment targets for public health policy is a necessary pragmatic process, especially when fighting endemic diseases.

Acknowledgements

The authors thank the villagers who agreed to participate in the study. Gratitude is extended to support staff of the involved institutions: the Direction de la Lutte contre les maladies transmissibles of the Malagasy Health Ministry, the French Institut Pasteur of Madagascar, and also to following researchers and technicians who participated in the study: Laurent Brutus, Virginie Hanitrasoamampionona, Georges Hébrard, Jacques Prod'Hon, Vohangy Ravaoalimalala, Pascaline Ravoniarimbinina, and Hélène Razanatsoarilala. The RAMSE program was financed by the French Institut de Recherche pour le Développement.

KEY MESSAGES

- Studies are used to explore the influence of either individual or collective factors on health outcomes but more analyses simultaneously focusing on the different levels of the social organization would substantially support the epidemiological approach to diseases.
- The failure to explicitly model the structure of such complex data is to ignore information about variability that, potentially, is as important as knowledge of the average effects.
- Multilevel modelling offers the opportunity to determine the relative impact of each level of organization on the variability and to identify the factors at each level that are associated with that level's impact.

References

[1] Susser M, Susser E. Choosing a future for epidemiology. I, Eras and paradigms. *Am J Public Health* 1996;**86:**668–73.

[2] Pearce N. Traditional epidemiology, modern epidemiology and public health. *Am J Public Health* 1996;**86:**678–83.

[3] Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *Am J Public Health* 1998;**88:**216–22.

[4] Humphreys K, Carr-Hill R. Area variations in health outcomes: Artefact or Ecology. *Int J Epidemiol* 1991;**20:**251–58.

[5] Susser M, Susser E. Choosing a Future for epidemiology. II, From black box to Chinese boxes and eco-epidemiology. *Am J Public Health* 1996;**86:**674–77.

[6] Verheij RA. Explaining urban-rural variations in health: a review of interactions between individual and environment. *Soc Sci Med* 1996;**42:**923–35.

[7] Boyle MH, Willms JD. Place effects for areas defined by administrative boundaries. *Am J Epidemiol* 1999;**149:**577–85.

[8] Blakely TA, Woodward AJ. Ecological effects in multi-level studies. *J Epidemiol Community Health* 2000;**54:**367–74.

[9] Diez-Roux AV. Investigating neighborhood and area effect on health. *Am J Public Health* 2001;**91:**1783–89.

[10] Roberts EM. Neighborhood social environments and the distribution of low birthweight in Chicago. *Am J Public Health* 1997;**87:**597–603.

[11] Davey Smith G, Hart C, Blane D, Gillis C, Hawthorne V. Lifetime socioeconomic position and mortality: prospective observational study. *BMJ* 1997;**314:**547–52.

[12] Moore DA, Carpenter TE. Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiol Rev* 1999;**21:**143–61.

[13] Beck LR, Rodriguez MH, Dister SW *et al*. Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *Am J Trop Med Hyg* 1994;**51:**271–80.

[14] Hay SI, Snow RW, Rogers DJ. Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. *Trans R Soc Trop Med Hyg* 1998;**92:**12–20.

[15] Kleinschmidt I, Bagayoko M, Clarke GP, Craig M, Le Sueur D. A spatial statistical approach to malaria mapping. *Int J Epidemiol* 2000;**29:**355–61.

[16] Kleinschmidt I, Sharp BL, Clarke GP, Curtis B. Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa. *Am J Epidemiol* 2001;**153:**1213–21.

[17] Duncan C, Jones K, Moon G. Context, composition and heterogeneity: using multilevel models in health research. *Soc Sci Med* 1998;**46:**97–117.

[18] Liang K, Zeger S. Regression analysis for correlated data. *Annu Rev Public Health* 1993;**14:**43–68.

[19] Morisson DA. Further difficulties with multifactorial analysis of variance: Random and nested factors and independence of data. *Infect Gen Evol* 2002;**2:**149–52.

[20] Diez-Roux AV. A glossary for multilevel analysis. *J Epidemiol Community Health* 2002;**56**:588–94.

[21] Bosma H, Van De Mheen HD, Borsboom G, Mackenbach JP. Neighborhood socioeconomic status and all-cause mortality. *Am J Epidemiol* 2001;**153:**363–71.

[22] Shouls S, Congdon P, Curtis S. Modelling inequality in reported long term illness in UK: combining individual and area characteristics. *J Epidemiol Community Health* 1996;**50:**366–76.

[23] Pickett KE, Pearl M. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *J Epidemiol Community Health* 2001;**55:**111–22.

[24] Johnstone FD, Raab GM, Hamilton BA. The effect of human immunodeficiency virus infection and drug use on birth characteristics. *Obstet Gynecol* 1999;**88:**321–26.

[25] Roepstorff A, Nilsson O, Ocallaghan CJ *et al*. Intestinal parasites in swine in the Nordic countries: multilevel modelling of *Ascaris suum* infections in relation to production factors. *Parasitology* 1999;**11:**521–34.

[26] Blanchy S, Rakotonjanabelo A, Ranaivoson G, Rajaonarivelo E. Epidémiologie du paludisme sur les hautes terres malgaches depuis 1878. *Cah Sante* 1993;**3:**155–61.

[27] Mouchet J, Blanchy S, Rakotonjanabelo A *et al*. Epidemiological stratification of malaria in Madagascar. *Arch Inst Pasteur Madagascar* 1993;**60:**50–59.

[28] Lepers JP, Deloron P, Andriamangatiana-Rason MD, Ramanamirija JA, Coulanges P. Newly transmitted *Plasmodium falciparum* malaria in the central Highlands of Madagascar: assessment of its clinical impact in a rural community. *Bull World Health Organ* 1990; **68:**217–22.

[29] Goldstein H. *Multilevel Statistical Models. 2nd Edn*. London: Edward Arnold, 1995.

[30] Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. *J Roy Stat Soc* 1996;**159:**505–13.

[31] Greenland S. Principles of multilevel modelling. *Int J Epidemiol* 2000;**29:**158–67.

[32] Rasbash J, Browne W, Goldstein H *et al. A user's guide to MlwiN. 2nd Edn*. London: Institute of Education, 2000.

[33] Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998;**17:**261–91.

[34] Koopman JS, Prevots DR, Vaca-Marin MA *et al*. Determinants and predictors of dengue infection in Mexico. *Am J Epidemiol* 1991;**133:**1168–78.

[35] Susser M. The logic in ecological: II. The logic of design. *Am J Public Health* 1994;**84:**830–35.

[36] Ghebreyesus TA, Haile M, Witten KH *et al*. Household risk factors for malaria among children in the Ethiopian highlands. *Trans Roy Soc Trop Med Hyg* 2000;**94:**17–21.

[37] Baird JK. Host age as a determinant of naturally acquired immunity to *Plasmodium falciparum*. *Parasitol Today* 1995;**11:**105–11.

[38] Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-average and subject-specific approaches for analysing repeated binary outcomes. *Am J Epidemiol* 1998;**147:**694–703.

[39] Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Stat Med* 2000;**19:**2675–88.

[40] Goldstein H, Brown W, Rasbach J. Multilevel modelling of medical data. *Stat Med* 2002;**21:**3291–315.

[41] Zhou XZ, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. *Am Stat* 1999;**53:**282–90.

[42] Sullivan LM, Dukes KA, Losina E. Tutorials in biostatistics: an introduction to hierarchical linear modelling. *Stat Med* 1999;**18:**855–88.

[43] Multilevel Project. Software reviews of multilevel analysis packages. http://multilevel.ioe.ac.uk/softrev/index.html