Research Article

Statistics in Medicine

Received 10 September 2012, Accepted 8 January 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5752

Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models

Yehenew Getachew,[a,e] Paul Janssen,[b] Delenasaw Yewhalaw,[c] Niko Speybroeck[d] and Luc Duchateau[e*†]

To study the effect of a mega hydropower dam in southwest Ethiopia on malaria incidence, we have set up a longitudinal study. To gain insight in temporal and spatial aspects, that is, in time (period = year–season combination) and location (village), we need models that account for these effects. The frailty model with periodwise constant baseline hazard (a constant value for each period) and a frailty term that models the clustering in villages provides an appropriate tool for the analysis of such incidence data. Count data can be obtained by aggregating for each period events at the village level. The mixed Poisson regression model can be used to model the count data. We show the similarities between the two models. The risk factor in both models is the distance to the dam, and we study the effect of the risk factor on malaria incidence. In the frailty model, each subject has its own risk factor, whereas in the Poisson regression model, we also need to average the risk factors of all subjects contributing to a particular count. The power loss caused by using village averaged distance instead of individual distance is studied and quantified. The loss in the malaria data example is rather small. In such a setting, it might be advantageous to use less labor-intensive sampling schemes than the weekly individual follow-up scheme used in this study; the proposed alternative sampling schemes might also avoid community fatigue, a typical problem in such research projects. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: mixed Poisson regression; periodwise constant hazard; frailty model; power; malaria incidence

1. Introduction

To study the temporal (year–season combination) and spatial (village as a cluster) dynamics of malaria occurrence, malaria incidence data are required. The sampling plan of such a malaria incidence study can vary substantially in terms of sampling frequency. The malaria incidence data, collected in the present study, sit at one extreme of the sampling spectrum: children are followed up for malaria on a weekly basis, so that these data can be regarded as actual event times, up to an interval censoring period of 7 days, which is very small relative to the length of study period. Such an intensive sampling scheme, however, is costly not only in terms of logistics and budget but also in terms of community fatigue resulting from the intensive contribution to the research project. The other extreme of the spectrum is a cross-sectional study with the malaria status assessed only once at a particular time. This type of data collection allows to estimate prevalence but not incidence; moreover, it does not allow to investigate infectious disease dynamics. An essential question addressed in the paper is whether the intensive sampling scheme used in the current malaria incidence study is really needed and whether less demanding sampling schemes can be as effective.

The frailty model [1] can be used to analyse the clustered time to malaria data. A frailty model with piecewise constant baseline hazard (for each period (year–season combination) the baseline hazard takes

[a] *Department of Horticulture and Plant Sciences, Jimma University, Jimma, Ethiopia*
[b] *Center for Statistics, Hasselt University, Diepenbeek, Belgium*
[c] *Department of Biology, Jimma University, Jimma, Ethiopia*
[d] *Public Health School, Université Catholique de Louvain, Brussels, Belgium*
[e] *Department of Comparative Physiology and Biometrics, University of Ghent, Ghent, Belgium*
**Correspondence to: Luc Duchateau, Department of Comparative Physiology and Biometrics, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium.*
[†] *E-mail: luc.duchateau@ugent.be*

a constant value), a frailty term that accounts for the clustering of the data at the village level and – for each subject – the distance to the dam as risk factor provides an appropriate inferential tool for the analysis of the incidence data. By aggregating, for each period, events at the village level, count data are obtained, and mixed Poisson regression models can be used for the analysis. The frailty model and mixed Poisson regression model are tightly connected. In the context of the hierarchical Cox model, Do Ha and Lee [2] and Ma *et al.* [3] showed that their likelihoods correspond.

In this paper, we show that such likelihood correspondence also holds for parametric frailty models with periodwise constant baseline hazard and mixed Poisson regression models. Of course, mixed Poisson regression modelling is not based on individual risk factors; it needs an aggregated risk factor, that is, a risk factor that is the average of the risk factors of all the subjects contributing to a particular count.

Interest in this malaria study is mainly in the effect of the dam on malaria incidence. The dam, with its vast water surface, is an excellent breeding place for the *Anopheles* mosquito, which is the vector of the *Plasmodium* parasite that causes malaria. The expectation is therefore that children living close to the dam are more at risk for malaria.

For the time to malaria data, each subject has its own specific distance to the dam. For the Poisson regression, however, because events are aggregated according to village and period, the risk factor used for a particular count consists of the mean of the subject distances of a village to the dam. The loss of power due to the aggregation of the risk factor is studied through simulation. The important insight gained from the simulation is that less-intensive sampling schemes reach almost the same efficiency as the longitudinal cohort sampling scheme described earlier.

We provide details on the time-to-event data and their aggregated counts in Section 2. We discuss the marginal and conditional Poisson and hazard models in Section 3, where we also demonstrate the equivalence of the two models. We show the analysis results in Section 4, followed by a simulation study in Section 5 to compare the power of the mixed Poisson regression model and the frailty model. The conclusions are in Section 6.

2. Individual and aggregated incidence data

To assess the effect of a mega hydropower dam in southwest Ethiopia on malaria incidence in children younger than 10 years, a longitudinal cohort study was set up. A total of 2082 children younger than 10 years from 16 villages located at different distances to the dam shore, with a range of 0.26 to 9.05 km, were enrolled and followed up from July 2008 to June 2010 at weekly intervals based on house-to-house visits. Details of the study design and study population are described elsewhere [4].

We investigated the effect of distance to the dam on malaria incidence. We cluster the children within villages. Within each village i $(i = 1, \ldots, 16)$, we follow up a number of children j $(j = 1, \ldots, n_i)$ for malaria (Figure 1). The number of children per village is on average 130 and does not differ much from one village to another. We observe y_{ij}, the minimum of the censoring time c_{ij}, and the event time t_{ij} and $\delta_{ij} = I(t_{ij} \leqslant c_{ij})$, the censoring indicator. We study the effect of x_{ij}, the distance to the dam, on the (possibly censored) event time. The y_{ij}'s are used for survival modelling. In Poisson regression models, however, we need to aggregate these individual event times, both in terms of time and space. The time axis is split into six periods (three seasons per year and two study years).

Period k starts at time r_{k-1} and ends at time r_k. As children are clustered within villages, the most obvious spatial aggregation level is the village. Aggregating event times for each period–village combination leads to a small set of summary statistics. The first summary statistic for village i is the number of events in period k, $d_{ik} = \sum_{j=1}^{n_i} \delta_{ij} I(r_{k-1} < y_{ij} \leqslant r_k)$. The second summary statistic is the total time at risk for village i in period k, $a_{ik} = \sum_{j=1}^{n_i} (\min(y_{ij}, r_k) - r_{k-1}) I(y_{ij} > r_{k-1})$. Counts d_{ik} are assumed to be independent between villages; the proposed model will accommodate for dependence between counts in different periods in the same village.

The summary statistics used in the Poisson regression correspond to (d_{ik}, a_{ik}), $i = 1, \ldots, 16$, $k = 1, \ldots, 6$.

3. Count and survival models

In Section 3.1, we consider marginal models. We study in Section 3.2 conditional models, that is, models where we add random effects.

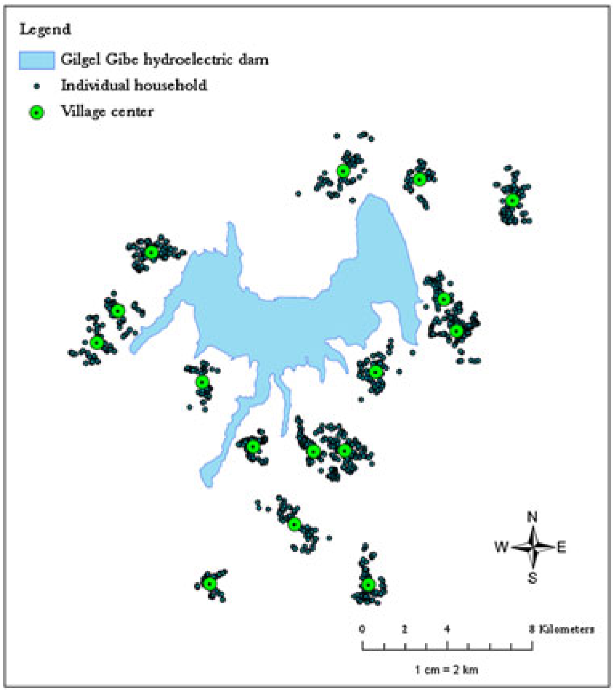

Figure 1. Map of the study area showing the distribution of Gilgel Gibe hydroelectric dam reservoir, study villages and children in southwest Ethiopia.

3.1. *The marginal model*

In the count regression model, we use, for $k = 1, \dots, 6$ and $i = 1, \dots, 16$, d_{ik} as response variable and the corresponding total time at risk, a_{ik}, as fixed offset variable. It is obvious that the individual distance from the child to the dam, the actual risk factor, cannot be used for these aggregated data. We rather use the average distance of the village to the dam, denoted by $\bar{x}_{i.} = \left(\sum_{j=1}^{n_i} x_{ij}\right)/n_i$ for village i. Furthermore, we have to adjust for period as a large part of the variation in malaria incidence is due to yearly and seasonal variation. We therefore write the Poisson regression model in terms of the expected number of events $\xi_{ik} = \mathrm{E}(d_{ik})$ as

$$\log(\xi_{ik}) = \log(a_{ik}) + \eta_{ik} \tag{1}$$

where

$$\eta_{ik} = \beta_0 + \beta_y x_{yk} + \beta_{s2} x_{s2,k} + \beta_{s3} x_{s3,k} + \beta_d \bar{x}_{i.}$$

with

$$x_{yk} = \begin{cases} 1 & k > 3 \\ 0 & k \leqslant 3 \end{cases}$$

$$x_{s2,k} = \begin{cases} 1 & k = 2, 5 \\ 0 & \text{otherwise} \end{cases}$$

$$x_{s3,k} = \begin{cases} 1 & k = 3, 6 \\ 0 & \text{otherwise} \end{cases}$$

and β_{s2}, β_{s3}, β_y and β_d the effect of the second season, the third season, the second year and the distance, respectively.

The data are connected to the model by the distributional assumption $d_{ik} \sim \mathrm{Poisson}(\xi_{ik})$.

The parametric survival model, on the other hand, is based on the individual data. We propose the hazard model

$$h_{ij}(t) = \sum_{k=1}^{6} \exp(\lambda_{ijk}) I(r_{k-1} < t \leqslant r_k) \tag{2}$$

where

$$\lambda_{ijk} = \alpha_0 + \alpha_y x_{yk} + \alpha_{s2} x_{s2,k} + \alpha_{s3} x_{s3,k} + \alpha_d x_{ij}$$

or

$$\exp(\lambda_{ijk}) = \exp\left(\alpha_0 + \alpha_y x_{yk} + \alpha_{s2} x_{s2,k} + \alpha_{s3} x_{s3,k}\right) \exp\left(\alpha_d x_{ij}\right)$$

with α_{s2}, α_{s3}, α_y and α_d the effect of the second season, the third season, the second year and the distance, respectively. We consider the first factor as the baseline hazard, which is constant within each of the six periods. This piecewise constant baseline hazard is different from the usual piecewise constant hazard model in the sense that the constant hazards are functions of the season and year effect, which allows evaluation of such effects on the baseline hazard. More flexible baseline hazards can be used on the basis of splines [5].

None of the two aforementioned models takes into account the dependence in the data. There is on the one hand the dependence in time between the counts measured at different periods in the same village, that is, between $d_{i1}, \ldots, d_{i6}$. On the other hand, there is the dependence in space between the times measured for different children in the same village, that is, between $y_{i1}, \ldots, y_{in_i}$. Although the estimates $\hat{\beta}_d$ and $\hat{\alpha}_d$ from these independence working models are consistent estimators for the population-based parameters, their standard errors are not. Sandwich estimators that cope with the dependence in the data should be used to adjust the standard error [6,7].

3.2. *The conditional model*

An alternative to cope with the dependence in the observations is the conditional model. A nice feature of the conditional model is that it provides information on the strength of the dependence.

The mixed Poisson regression model is given by

$$\log(\xi_{ik}) = \log(a_{ik}) + \eta_{ik} + w_i \tag{3}$$

with w_i, the random effect, normally distributed with mean zero and variance σ_p^2.

The mixed survival model, the frailty model [1], is given by

$$h_{ij}(t) = \sum_{k=1}^{6} \exp(\lambda_{ijk} + u_i) I(r_{k-1} < t \leqslant r_k) \tag{4}$$

with u_i assumed to be normally distributed with mean zero and variance σ_f^2, which describes the heterogeneity between clusters (villages).

The presence of the random effect in (3) makes $\log(\xi)$ a random variable. Insight in the effect of the heterogeneity between clusters (villages) on the 'expected' count ξ can be obtained by looking at its density

$$f_\xi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(-\frac{(\log(x/a) - \eta)^2}{2\sigma^2}\right) \tag{5}$$

For frailty models, we can translate the heterogeneity at the level of the hazard function to heterogeneity in terms of the median malaria event time or in terms of the percentage of subjects having had malaria at a particular time over the different villages; see [8] for a detailed discussion.

3.3. *Equivalence of the count and survival model*

Although the model specifications of the mixed Poisson regression model (3) and the frailty model (4) are quite different, we show that the two modelling approaches give the same parameter estimates, apart from the fact that the mixed Poisson regression model uses mean risk factors, that is, mean of the children distances to the dam, whereas the frailty model uses individual risk factors, that is, individual child distance to the dam.

The conditional loglikelihood contribution of the ith village in the kth period is given by

$$\ell_{pik} = d_{ik} \log(\xi_{ik}) - \xi_{ik} - \log(d_{ik}!)$$

Dropping the last term, a constant, and replacing ξ_{ik} using (3), we obtain

$$\ell p_{ik} = d_{ik} \log(a_{ik}) + d_{ik}(\eta_{ik} + w_i) - a_{ik} \exp(\eta_{ik} + w_i) \tag{6}$$

On the other hand, we consider the frailty model replacing the individual child distance to the dam x_{ij} with the village mean $\bar{x}_{i.}$, leading to model

$$h_{ij}(t) = \sum_{k=1}^{6} \exp\left(\lambda_{ik} + u_i\right) I\left(r_{k-1} < t \leqslant r_k\right) \tag{7}$$

with

$$\lambda_{ik} = \alpha_0 + \alpha_y x_{yk} + \alpha_{s2} x_{s2,k} + \alpha_{s3} x_{s3,k} + \alpha_d \bar{x}_{i.}$$

Using (7), the cumulative hazard for $r_{k-1} < t < r_k$ is given by

$$\begin{aligned} H_{ij}(t) &= (r_1 - r_0)\exp(\lambda_{i1} + u_i) \\ &\quad + (r_2 - r_1)\exp(\lambda_{i2} + u_i) \\ &\quad \vdots \\ &\quad + (t - r_{k-1})\exp(\lambda_{ik} + u_i) \\ &= \sum_{k=1}^{6} I(t > r_{k-1})\left(\min(r_k, t) - r_{k-1}\right)\exp(\lambda_{ik} + u_i) \end{aligned}$$

from which follows

$$\log S_{ij}(t) = -H_{ij}(t) = -\sum_{k=1}^{6} I(t > r_{k-1})(\min(r_k, t) - r_{k-1})\exp(\lambda_{ik} + u_i) \tag{8}$$

The conditional loglikelihood contribution of the jth child in the ith village is generally given by

$$\ell s_{ij} = \delta_{ij} \log(h_{ij}(y_{ij})) + \log(S_{ij}(y_{ij}))$$

and using (7) and (8), we have

$$\ell s_{ij} = \sum_{k=1}^{6} \delta_{ijk}(\lambda_{ik} + u_i) - I(y_{ij} > r_{k-1})(\min(r_k, y_{ij}) - r_{k-1})\exp(\lambda_{ik} + u_i)$$

with δ_{ijk} denoting whether an event takes place ($\delta_{ijk} = 1$) or not ($\delta_{ijk} = 0$) in period k for the particular child.

Now summing over all children in the village and splitting up the sum over the six different periods, we obtain

$$\ell s_{ik} = d_{ik}(\lambda_{ik} + u_i) - a_{ik}\exp(\lambda_{ik} + u_i) \tag{9}$$

It is now easy to see that (6) and (9) are the same loglikelihood expressions, and parameter estimates will thus be exactly the same.

4. The malaria incidence data analysis

We analysed the malaria incidence data by using both the marginal and conditional models and both the Poisson regression and hazard models. We can obtain the marginal model parameter estimates by fitting the model to the data not taking into consideration the dependence structure in the data and next replace the standard error by its robust sandwich estimator. For the conditional model, we apply Gaussian quadrature to numerically integrate out the normally distributed random effects, after which the resulting likelihood expression can be maximised [9]. We based the model fitting on the SAS nlmixed procedure (Appendix A) and present the results in Table I.

Table I. Parameter estimates (standard error) for the malaria incidence data from the marginal and conditional Poisson regression and hazard models.

Parameter	Marginal Poisson	Marginal hazard	Conditional Poisson	Conditional hazard
β_0/α_0	−7.302 (0.164)	−7.269 (0.152)	−7.389 (0.244)	−7.167 (0.255)
β_{s2}/α_{s2}	−1.543 (0.183)	−1.543 (0.183)	−1.524 (0.136)	−1.524 (0.136)
β_{s3}/α_{s3}	−0.826 (0.104)	−0.827 (0.103)	−0.803 (0.119)	−0.802 (0.119)
β_y/α_y	−0.191 (0.081)	−0.192 (0.080)	−0.154 (0.088)	−0.154 (0.088)
β_d/α_d	0.069 (0.053)	0.056 (0.056)	0.054 (0.074)	−0.035 (0.057)
σ^2/θ			0.310 (0.129)	0.349 (0.148)

Hazard models are based on the individual child distance to the dam, Poisson models on the mean of the children distances of the village to the dam. The parameter β_0 corresponds to the logarithm of the expected number of events over the at-risk time in the first period (year 1 and season 1) in the Poisson model, with β_{s2}, β_{s3}, β_y and β_d the effect of the second season, the third season, the second year and the distance, respectively. The parameter α_0 corresponds to the baseline hazard in the first period (year 1 and season 1) in the hazard model, with α_{s2}, α_{s3}, α_y and α_d the effect of the second season, the third season, the second year and the distance, respectively. The parameters σ^2 and θ correspond to the variance of the random effects in the conditional Poisson and frailty models, respectively.

Although in none of the models a significant effect of distance was found, it is to be noted that three of four models predict increasing incidence with increasing distance to the dam. Only the frailty model predicts a decreasing incidence with increasing distance to the dam. It is noteworthy that replacing the individual distance with the mean distance from the village will lead to the same results for the mixed Poisson regression model and the frailty model. Therefore, the fact that the effect of the distance changes direction is entirely due to the use of individual distances. This point will be taken up further in the discussion.

We can further make use of (5) to have some insight in the data heterogeneity. As the outcome of interest is the expected count, the at-risk time of the different villages needs to be set at a same value a. We determine the fixed offset term a for the first season in the first year, the period with the highest malaria incidence, as follows. The total time at risk over all villages in that period equals 289 841 days, corresponding to an average time at risk per village equal to 18 115 days, which is the value used for a. We further set $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_d \times \text{distance} = -7.389 + 0.054 \times \text{distance}$ and $\sigma_p^2 = 0.310$. We depict the density functions for three different distances to the dam (1, 4 and 8 km) in Figure 2. The average of the expected count (the mean of the density) is increasing with increasing distance to the dam, with highest average expected count at 1 km. When considering the density function of the expected count over villages at equal distance to the dam, for instance at 1 km, represented by the solid line in Figure 2, the variance of that density function is much larger than the variation between the average expected count at the different distances.

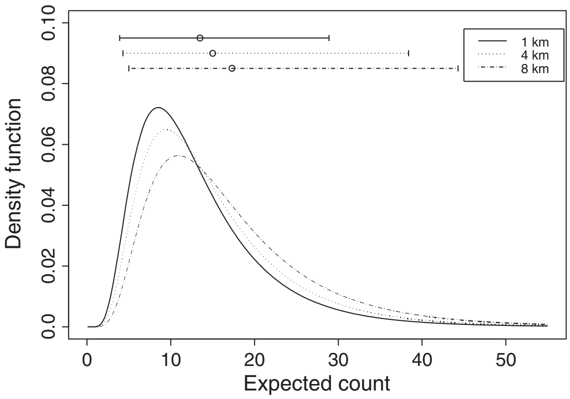

Figure 2. Density functions of the expected count in villages as a function of distance to the dam. The mean (circle) with 95% interval for the three distances is given in the top panel.

5. The power of count and survival models

The main difference between the mixed Poisson regression model and the periodwise constant hazard frailty model is the ability to make use of aggregated versus individual risk factors. Therefore, we study the effect of the individual versus the averaged risk factor on the size and the power of the test for that particular risk factor.

We simulate data from the periodwise constant hazard frailty model using parameters as estimated from the model and presented in Table I. We use the individual distances observed in the data as risk factors and generate a total of 5000 data sets for each value of α_d that is assessed. The simulation results are summarised by the median, 5th and 95th quantile of the 5000 generated results for $\hat{\alpha}_d$ and its standard error, and further the coverage of the 95% confidence interval in Table II. When generating data under the assumption $\alpha_d = 0$, the coverage equals 0.91 and 0.93 for the mixed Poisson and frailty model, respectively, converting to test sizes equal to 0.09 and 0.07. When generating data under the assumption $\alpha_d > 0$, the observed bias is rather small for both models, but the coverage of the frailty model, equal to 0.92, is always slightly better than that of the mixed Poisson regression model, which is between 0.88 and 0.90.

The power is defined as the number of times the confidence interval of α_d does not contain zero when generating data under a specific alternative assumption $\alpha_d \neq 0$. We present the simulation results on the power in Figure 4. Within the range 0.08–0.15 for α_d, the largest power gain for the frailty model is observed, with a maximum for $\alpha_d = 0.12$ corresponding to a difference in power equal to 0.123.

6. Discussion

The frailty model with village as frailty term and baseline hazard constant within period is equivalent to the mixed Poisson regression model with village as random effect when counts are aggregated within period and village, whenever the covariate information is constant within a village. In our data setting, however, the covariate of interest, that is, distance to the dam, is changing within village. This has important consequences on the analyses results of the two models and also on the power of particular sampling schemes.

Although none of the two conditional models leads to a significant effect of the distance to the dam on malaria incidence, it is remarkable that the effect changes direction from one model to the other, merely because the frailty model uses individual distance compared with the mixed Poisson regression model that is based on the average distance of the village. This difference can be explained as follows. Consider the best linear unbiased predictors of the frailties u_i as a function of the distance to the dam in Figure 3. There seems to be a positive relationship between the predictors $\hat{u}_i$ and the distance to the dam, which corresponds to the positive relationship observed in the frailty model with average distance. The parameter α_d in the frailty model with individual distance describes the effect of distance over and

Table II. Simulation results for the malaria incidence data comparing the mixed Poisson regression and hazard models, assuming different values for α_d.

Parameter value	Model	Median estimate (P5, P95)	Median se(β) (P5, P95)	Coverage
0.00	Poisson	0.000 (−0.087, 0.079)	0.045 (0.031, 0.061)	0.91
	Hazard	0.000 (−0.071, 0.068)	0.040 (0.029, 0.049)	0.93
0.05	Poisson	0.049 (−0.034, 0.130)	0.044 (0.031, 0.060)	0.90
	Hazard	0.050 (−0.019, 0.118)	0.039 (0.029, 0.048)	0.92
0.10	Poisson	0.098 (0.012, 0.176)	0.044 (0.030, 0.059)	0.88
	Hazard	0.099 (0.033, 0.164)	0.038 (0.028, 0.046)	0.92
0.15	Poisson	0.149 (0.065, 0.227)	0.043 (0.029, 0.059)	0.88
	Hazard	0.151 (0.084, 0.213)	0.037 (0.027, 0.045)	0.92
0.20	Poisson	0.198 (0.117, 0.274)	0.042 (0.029, 0.058)	0.89
	Hazard	0.200 (0.137, 0.262)	0.036 (0.027, 0.043)	0.92

The frailty model is based on the individual child distance to the dam, the mixed Poisson regression models on the mean of the children distances of the village to the dam. The median of the estimate and the standard error (with 5th (P5) and 95th (P95) percentile) is reported for $\hat{\beta}_d$ (Poisson) and $\hat{\alpha}_d$ (Hazard), together with the coverage (aiming at a 95% confidence interval).

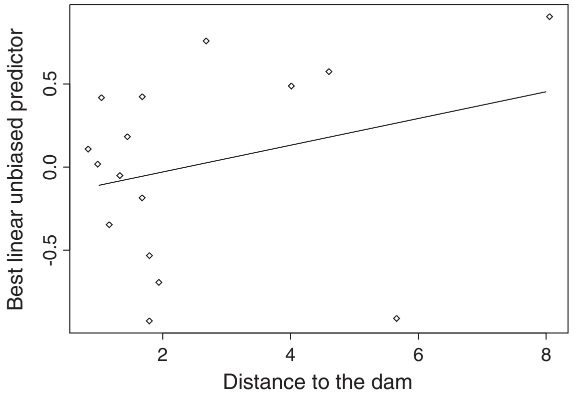

Figure 3. Best linear unbiased predictors (BLUPs) of the village effects represented as a function of distance to the dam. The BLUPs are obtained from the frailty model with individual time to malaria data as response variable, individual distance as a fixed effects factor and village as a frailty term. The line corresponds to the regression of the BLUPs on the distance to the dam.

above the average distance effect of the village captured by the frailty term, whereas the parameter β_d in the mixed Poisson regression model describes the effect of distance between villages. It is also not surprising that the estimate of the parameter α_d in the marginal hazard model is close to both the conditional and marginal Poisson models, because the effect of the distance in the marginal hazard model is based on all subjects without taking into consideration the clustering and especially those children far away and close to the dam will have a large influence on the parameter estimate. From a practical point of view, the results from either the mixed Poisson regression model or the marginal hazard model seem to be more relevant, as interest is in the distance effect at the larger scale between villages.

Both models, the frailty model and the mixed Poisson regression model, lead to sizes above the significance level, although in the case of the frailty model, with a size of 7%, it is only slightly above the significance level of 5%. This is because there is no sufficient replication at the cluster level. On the basis of simulations (data not shown), doubling the number of clusters to 32 will bring the size down to 6.5% for the mixed Poisson regression model. We can conclude that 16 clusters is really at the borderline and preferably more clusters should be available.

The aggregation of the time-to-event data to count data reduces information, which has an effect on the power as demonstrated in Figure 4. Therefore, if the total time at risk is fixed and the power needs to be maximised, individual event times should be collected and analysed by the frailty model. The power reduction, however, is rather small. Under certain assumptions, this opens up opportunities for improved sampling schemes.

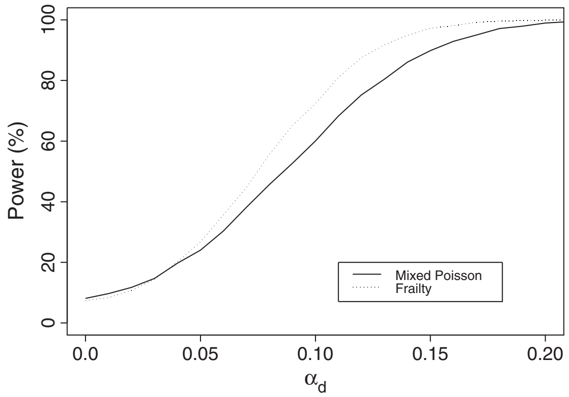

Figure 4. The power of the mixed Poisson regression and frailty models to test for the effect of the distance of the dam for different alternative values for that effect α_d.

Under the assumption of periodwise constant baseline hazard, event times can be aggregated to counts that can be modelled using the mixed Poisson regression model. It means that it is no longer necessary to follow up individuals as intensively as in the present study with continuous subject follow-up. It would be more appropriate to include a larger number of subjects over more villages that are followed up only for a brief period. The increase in the number of clusters and the number of subjects will make the conclusions more general on the one hand and will provide more replication at the level of the cluster, which will improve the size of the test. Another important advantage is the reduction of the risk of community fatigue; it is often observed that with long follow-up studies, study subjects adhere less and less to the protocols.

In a small simulation study (data not shown) based on the same setting as Section 5, we reduced the time at risk in each village to 50% of the time at risk in the original data set, which can be carried out in practice by either reducing the number of children in the village or reducing the time to follow-up of each individual child. This substantial reduction in time at risk in each village will reduce the power for $\alpha_d = 0.1$ from 59.8% to 57.4%, which corresponds to a negligible decrease in power for a substantial decrease in workload. This warrants further research to run such studies in a more efficient way.

The mixed Poisson regression seems to be a good alternative for the frailty model. Aggregation in periods, however, remains essential, as it has been shown that reducing time-to-event data to binary data (event or not in the whole follow-up period) reduces the power of the test substantially [10].

Appendix A. SAS program (nlmixed procedure) to fit mixed Poisson and frailty model

```
data malaria;
input cluster time cens dist;
cards;
1  230 1 3.45
1  720 0 3.67
...
16 465 1 7.88
;

proc sort data=malaria;by cluster;
proc means data=malaria noprint;var dist;output out=dist mean=distm;by cluster;

data quant_d;
qd0=0;qd1=150;qd2=270;qd3=365;qd4=515;qd5=635;qd6=730;
data all;set malaria;if _n_=1 then set quant_d;
data new; set all;
array quant_d{7} qd0 qd1 qd2 qd3 qd4 qd5 qd6;
array d{6} d1-d6;
array e{6} e1-e6;
do i=1 to 6;d{i}=0;e{i}=0;end;
do i=2 to 7;
 if time<=quant_d{i} then do;e{i-1}=(cens=1);d{i-1}=time-quant_d{i-1};i=7;end;
 else d{i-1}=quant_d{i}-quant_d{i-1};
 end;

data haz;merge new dist;by cluster;

/*Frailty model with individual distance (for mean distance replace dist with distm)*/
proc nlmixed data=haz;
parms l=1 s2=1 s3=1 y2=1 beta=1 sig2f=1;
bounds theta >= 0;

basehaz=exp(l*e1+(l+s2)*e2+(l+s3)*e3+(l+y2)*e4+(l+y2+s2)*e5+(l+y2+s3)*e6+beta*dist+nu);
cumhaz=(exp(l)*d1+exp(l+s2)*d2+exp(l+s3)*d3+exp(l+y2)*d4+exp(l+y2+s2)*d5+
        exp(l+y2+s3)*d6)*exp(beta*dist+nu);
loglik0=-cumhaz;
if cens=0 then loglik=loglik0;
if cens=1 then loglik=log(basehaz)+loglik0;
model time~general(lo lik);
random nu~normal(0,sig2f) subject=cluster;
run;
```

```
proc means data=new noprint;var d1 d2 d3 d4 d5 d6 e1 e2 e3 e4 e5 e6;
output out=datcount sum=d1 d2 d3 d4 d5 d6 e1 e2 e3 e4 e5 e6;by cluster;
data datcount (keep=cluster dist period year season y a);set datcount;
period=1;year=1;season=1;y=e1;a=d1;output;
period=2;year=1;season=2;y=e2;a=d2;output;
period=3;year=1;season=3;y=e3;a=d3;output;
period=4;year=2;season=1;y=e4;a=d4;output;
period=5;year=2;season=2;y=e5;a=d5;output;
period=6;year=2;season=3;y=e6;a=d6;output;
data datcount;merge datcount dist;by cluster;

data datcount (keep=cluster distm loga y season year season2 season3 year2);
set datcount;loga=log(a);season2=0;
if season=2 then season2=1;season3=0;if season=3 then season3=1;
year2=0;if year=2 then year2=1;

/*Mixed Poisson regression model with mean distance*/
proc nlmixed data=datcount;
parms logsigp 0 int 1 bs2 1 bs3 1 by2 1 betad 1;
eta=int+ loga+bs2*season2+bs3*season3+by2*year2+betad*distm+e;
lambda=exp(eta);
model y ~ poisson(lambda);
random e~normal(0,exp(2*logsigp)) subject=cluster;
run;
```

Acknowledgements

This research was supported by the Interuniversity Attraction Poles research network P7/StUDyS of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) and the Institutional University Cooperation Programme (Jimma University) of the Flemish Interuniversity Council (VLIR). The authors thank Saghir Bashir for the useful comments and the anonymous reviewers for their critical review that has improved the paper substantially.

References

1. Duchateau L, Janssen P. *The Frailty Model*. Springer: New York, 2008.
2. Do Ha I, Lee Y. Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics* 2003; **12**:663–681.
3. Ma RJ, Krewski D, Burnett RT. Random effects Cox models: a Poisson modelling approach. *Biometrika* 2003; **90**:157–169.
4. Yewhalaw D, Kassahun W, Woldemichael K, Tushune K, Sudaker S, Kaba D, Duchateau L, Van Bortel W, Speybroeck N. The influence of the Gilgel-Gibe hydroelectric dam in Ethiopia on caregivers' knowledge perceptions and health-seeking behavior towards childhood malaria. *Malaria Journal* 2010; **9**(47):1–11.
5. Nielsen J, Parner ET. Analyzing multivariate survival data using composite likelihood and flexible parametric modelling of the hazard functions. *Statistics in Medicine* 2010; **29**:2126–2136.
6. White D. Maximum likelihood estimation under mis-specified models. *Econometrica* 1982; **50**:1–26.
7. Zhang H, Yu Q, Feng C, Gunzler D, Wu P, Tu XM. A new look at the difference between the GEE and the GLMM when modeling longitudinal count responses. *Journal of Applied Statistics* 2012; **39**(9):2067–2079.
8. Duchateau L, Janssen P. Understanding heterogeneity in generalized mixed and frailty models. *American Statistician* 2005; **59**:143–146.
9. Liu L, Huang X. The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Statistics in Medicine* 2008; **27**:2665–2683.
10. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cut point. *Epidemiology* 1992; **3**:434–440.