STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:6489–6504
Published online 12 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3439

Frailty modelling for clustered recurrent incidence of diarrhoea

E. Elgmati[1, *, †], R. Fiaccone[2], R. Henderson[1] and M. Mohammadi[1]

[1] *School of Mathematics and Statistics, University of Newcastle upon Tyne, Tyne NE1 7RU, U.K.*
[2] *Departamento de Estatistica, Universidade Federal da Bahia, Campus Ondina 40170-290, Salvador-BA, Brazil*

SUMMARY

Recurrent incidence of infant diarrhoea is studied, using daily data collected in Salvador, Brazil, from 754 children over 455 days. Aalen's additive intensity model is taken as the basis of the modelling strategy and a frailty extension is proposed. The idea is to estimate the frailty dynamically as time proceeds and information accrues. This provides an alternative to the inclusion of dynamic covariates based on individual event patterns. Simulation results indicate good performance of the estimation methods. In our first analysis, there is no account taken in the natural clustering of the children into 21 different districts of the city. The model is therefore extended to incorporate the possibility of spatial or spatio-temporal clustering effects, possibly caused by unobserved environmental factors or infectivity. Significant frailty and significant clustering are both identified in the data. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: additive regression model; clustering; diarrhoea incidence; frailty; recurrent events

1. INTRODUCTION

This paper describes the analysis of data from a study carried out in northeast Brazil into the incidence of infant diarrhoea. The northeast region of Brazil, with a population of 40 million, is the poorest area of the country. Diarrhoea is the main cause of infant deaths, responsible for 13.6 per cent of deaths in children under 1 year old in 1995 [1].

The data to be considered are taken from a household survey carried out through home visits over 455 days from October 2000 to January 2002 [2]. One child aged under three years at entry was monitored from each household and in this paper we will concentrate on the 754 who were recruited at the beginning of the study and who had at least 90 days of follow-up. For each child we have records of incidence of episodes of diarrhoea, an episode being a sequence of days with

[*] Correspondence to: E. Elgmati, School of Mathematics and Statistics, University of Newcastle upon Tyne, Tyne NE1 7RU, U.K.
[†] E-mail: entisar.elgmati@newcastle.ac.uk

Contract/grant sponsor: National Counsel of Technological and Scientific Development—CNPq

Received 6 August 2008
Accepted 6 August 2008

Copyright © 2008 John Wiley & Sons, Ltd.

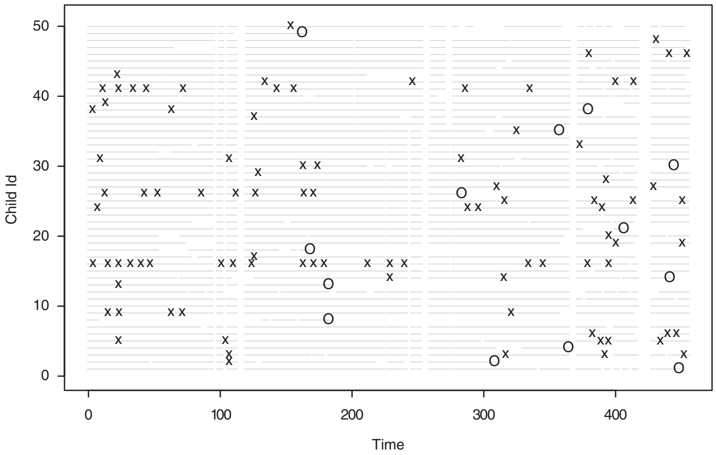

Figure 1. Incidence and dropout patterns for the first 50 children. Each line represents a child, with gaps where data are missing. Crosses represent new episodes and circles indicate dropouts.

diarrhoea until there have been at least three consecutive clear days. Onset of episodes forms a process of events in time and leads naturally to the use of event-history methodology in studying both changes in incidence over calendar time and the relationship between incidence and various socio-economic and demographic factors, particularly sanitation arrangements. Figure 1 illustrates the data for the first 50 children in the study. There is one horizontal line per child, which is missing on days when there is no information, and the crosses mark the onset of episodes of diarrhoea.

Borgan *et al.* [3] provide further information on these data, augmented by an additional 172 children who were late entrants to the study and who will not be considered here. Those authors describe the advantages of an approach based on Aalen's additive intensity model [4, 5]. The advantages include the facility to capture time-varying effects with time-varying covariates, very quick computation, and straightforward adaptation to individuals being at risk or otherwise at any given time. The latter is important for the diarrhoea incidence study, because children are not at risk of a new episode until the previous one is cleared. In addition, some of the children dropped out early from the study and although there is no evidence of this being informative [3], methods still need to be adaptive to changing risk sets. Note that under this approach we use calendar time as time scale. Another approach would be to analyse inter-event or gap times between episodes of diarrhoea as a renewal type process (e.g. [6]). However, we are specifically interested in changes in intensity over calendar time and also it is unlikely that the processes are stationary, hence our approach and choice of scale.

To allow for significant heterogeneity between individuals, as can be seen in Figure 1, Borgan *et al.* [3] included so-called dynamic covariates in their models. These are internal or endogenous covariates summarizing the previous history of episodes [7, 8]. Their inclusion allows the model to incorporate the intuitive possibility that a child with a history of frequent diarrhoea may have different susceptibilities to future episodes than another child who has been so far diarrhoea-free but is otherwise comparable. Another approach to the same problem is to include a *frailty* or random effect term in the model to describe the unexplained heterogeneity [9]. This leads to the first of our two main aims in this paper. Inclusion of frailty terms within a proportional intensity

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

model is by now standard [10], but frailty within the additive model structure is less easily handled, at least for estimation [11]. In Section 2 we describe the test and apply an iterative approach in which we alternate between a marginal analysis to estimate the frailty distribution and a conditional approach to estimate, at each timepoint, the additive model parameters.

Our second main aim is to examine the possibility of clustering of the incidence of diarrhoea, both in space and time. Borgan *et al.* treat children as being independent but acknowledge in discussion that since diarrhoea can be infectious there could be association between the incidence processes of children living close together. In Section 3 of this paper, we investigate this possibility after demonstrating that the sample consists of 21 clusters of between 23 and 47 children. To begin with, we assume that the clustering is spatial only, but in Section 4 we show that space–time clustering is indeed present in the data, as might be expected.

Notation, the standard Aalen additive regression model and a frailty extension are described in Section 2, all based on an assumption of independence between children. In Section 2.2, we describe and test an estimation procedure and we apply the method to the diarrhoea data in Section 2.3. In Section 3.1, we recognize possible non-independence and describe the partition of the 754 children into clusters. In Section 3.2, we examine a variety of methods for testing for differences in intensity patterns between clusters. In Section 3.3, we extend the frailty approach to allow for heterogeneity between, and within, clusters and in Section 3.4 we fit the frailty model to the clustered data. Section 4 has a generally similar structure to Section 3 but allows the possibility of space–time variation. The paper is concluded with a brief discussion in Section 5.

2. THE AALEN ADDITIVE INTENSITY MODEL WITH FRAILTY

2.1. Notation and model construction

To begin with, we assume independence between children. Throughout we consider dropout and periods of intermittent missingness to be non-informative, which is equivalent to independent censoring in event-history methodology. The study period is from time zero to time τ, which is 455 days for the application.

Our work is based on the Aalen additive model for events in time [4, 5]. The counting process $N_i(t)$ for episodes for child i with vector of covariates $x_i(t)=(1,x_{i1}(t),x_{i2}(t),\dots,x_{ip}(t))'$ is assumed to have intensity

$$\lambda_i(t|F_{t^-},x_i(t))=Y_i(t)\alpha(t|x_i(t))$$

where F_{t^-} is the history of events, censoring, and covariates prior to time t, and $Y_i(t)$ is an at-risk indicator with value one when individual i is susceptible to new episodes being observed, and zero otherwise. Since F_{t^-} includes covariates, formally there is no need to have $x_i(t)$ in the conditioning above. We have included it however to emphasize the elements of F_{t^-} included in the model. We assume that

$$\alpha(t|x_i(t))=\beta_0(t)+\beta_1(t)x_{i1}(t)+\cdots+\beta_p(t)x_{ip}(t)$$

where $\beta_0(t)$ is the baseline hazard and $\beta_j(t),\ j=1,\dots,p$, are covariate coefficients. Estimation [3–5] is usually by least squares or generalized least squares at each event time and inference is

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

directed towards the cumulative coefficients

$$B_j(t)=\int_0^t \beta_j(u)\,\mathrm{d}u$$

To allow for the possibility of heterogeneity between children not explained by covariates, we can introduce subject-specific multiplicative random effects or *frailties* into the model as:

$$\lambda_i(t|F_{t^-}, x_i(t), Z_i) = Z_i Y_i(t)\alpha(t|x_i(t)) \tag{1}$$

However, since Z_i is not observed we have to work with the marginal intensity:

$$\lambda_i(t|F_{t^-}, x_i(t)) = E\{Z_i|F_{t^-}\}Y_i(t)\alpha(t|x_i(t)) \tag{2}$$

If Z_i is assumed to follow a gamma distribution with unit mean and variance ξ then it is straightforward to show that

$$E\{Z_i|F_{t^-}\} = \frac{1+\xi N_i(t^-)}{1+\xi\Lambda_i(t)} \tag{3}$$

where

$$\Lambda_i(t)=\int_0^t Y_i(u)\alpha(u|x_i(u))\,\mathrm{d}u$$

2.2. Estimation

To estimate the parameters, we adopt an iterative approach based on estimating the frailty parameter ξ from the marginal distribution of total event counts, and the other parameters, given ξ, from (2) and (3). Since ξ is estimated from the total events there is mild conditioning on the future and formally the martingale structure, which underpins that the non-frailty inference for this class of model no longer holds. However, our experience in simulations is that there is no observable consequence of the breakdown and the method works well.

The iterative algorithm proceeds as follows:

1. Assume that $E\{Z_i(t)|F_{t^-}\}=1$ for all times and individuals. Collect these into vectors $\hat{Z}(t)$ at each t.
2. Take an initial value for ξ.
3. At each event time t form a design matrix $X(t)$ of covariates and an intercept column, and combine at-risk indicators $Y_i(t)$ into a vector $Y(t)$. Estimate the regression parameters through

$$X^*(t) = Y(t)\hat{Z}(t)X(t)$$

$$\hat{\boldsymbol{\beta}}(t)\,\mathrm{d}t = [X^*(t)'X^*(t)]^{-1}X^*(t)'\,\mathrm{d}N(t)$$

   where $\mathrm{d}N(t)$ is a vector of indicators for event experiences at time t.
4. Estimate ξ, for fixed $\boldsymbol{\beta}(t)$, by maximizing the marginal likelihood of total event counts. This is obtained from the negative binomial distribution arising from a gamma mixture of Poisson

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table I. Properties of $\hat{\xi}$ estimated using procedure in Section 2. Results based on 100 simulated samples.

Sample size	ξ	Mean ($\hat{\xi}$)	S.D. ($\hat{\xi}$)
100	0.5	0.493	0.082
	1.0	0.980	0.148
250	0.5	0.495	0.052
	1.0	1.014	0.096
500	0.5	0.497	0.038
	1.0	1.012	0.070
1000	0.5	0.495	0.026
	1.0	1.012	0.048

random variables. The contribution of child i is

$$P(N_i(\tau)=n)=\frac{1}{n!}\frac{\Gamma\left(n+\frac{1}{\xi}\right)}{\Gamma(1/\xi)}\frac{[\xi\Lambda_i(\tau)]^n}{[1+\xi\Lambda_i(\tau)]^{n+1/\xi}}$$

5. Update the $\hat{Z}(t)$ using (3).
6. Repeat steps 3–5 until convergence.

Table I summarizes some of our simulation results. We took $\tau=100$, two time-constant binary covariates with distribution $P(x_i=0)=P(x_i=1)=\frac{1}{2}$, $\beta_0(t)=0.05$, and $\beta_1(t)=\beta_2(t)=0.1$. The number of events per individual is averaged 15 and ranged from 1 to 35. The mean and standard deviation of estimates of ξ are given for a variety of sample sizes and true values of ξ, in each case based on 100 simulated samples. Results are good. Similarly encouraging results were obtained for the estimation of $B_j(t)$.

2.3. Aalen model with frailty applied to diarrhoea data

The frailty variance was estimated as $\xi=0.81$ (bootstrap SE 0.097 from 100 samples), indicating considerable heterogeneity in the data. Table II summarizes results of testing for covariate effects. It is straightforward to adapt to the frailty model a statistic U_j developed for the Aalen model for testing $H_0\colon \beta_j(t)=0$ for all $t\in(0,\tau)$. The test statistic U_j is formed as $\int L_j(t)\,\mathrm{d}\hat{B}_j(t)$, where $L_j(t)$ is a weight function. Following Aalen we took $L_j(t)$ equal to $1/\{(X^*(t)'X^*(t)\}_{jj}^{-1}$, in analogy with linear regression, to give more weight to the more precisely estimated regions. If $L_j(t)$ is predictable then standard martingale results can be used to provide concise expressions for the variance of the U_j. In our case the expected frailties are, given ξ, functions of history only and hence $L_j(t)$ is predictable. As stated, estimation of ξ formally causes dependence on the future, but our experience is that this has no noticeable effect. Further information on inference for the additive model is available in [3], where the covariates for the diarrhoea data are also described. Results are generally similar to those obtained by [3], who used a dynamic covariate (the average number of events per unit time) term in the model to describe heterogeneity rather than a frailty approach. Incidence of diarrhoea is affected by having more than three people per bedroom, living in an area with open sewerage or rain affected accommodation, and having a

Table II. Diarrhoea data: tests for covariate effects under Aalen's additive model with frailty.

β	Test U	SE(U)	U/SE	p-Value
Low socio-economic status	−12.80	26.46	−0.48	0.628
Other children ⩽5 years	0.98	29.63	0.03	0.973
More than three people/bedroom	92.68	25.79	3.59	<0.001
Poor street quality	−11.27	24.53	−0.46	0.646
Contaminated water source	−15.09	25.12	−0.60	0.548
Contaminated water storage	32.54	26.31	1.24	0.216
Standing water	6.23	24.07	0.26	0.796
Open sewerage	115.89	22.38	5.18	<0.001
Rain affected accommodation	62.25	26.45	2.35	0.019
Mother <25 years	77.12	31.09	2.48	0.013
Gender	6.45	31.28	0.21	0.837
Starting age (months)	−8596.34	576.39	−14.91	<0.001

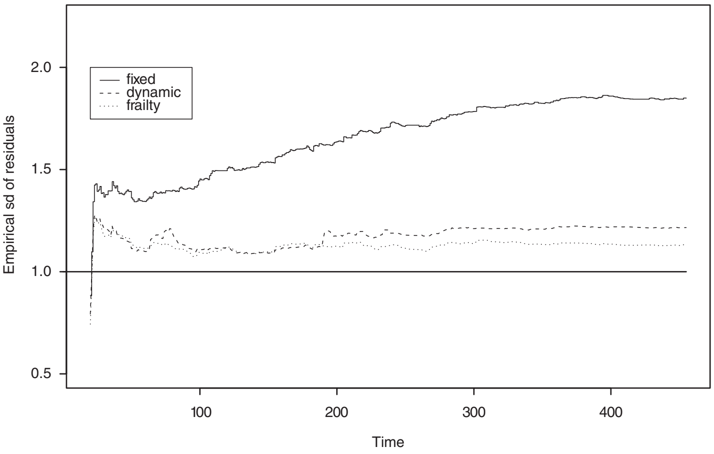

Figure 2. Empirical standard deviations of standardized martingale residual processes: solid line, only fixed covariates x included in the model; dashed line, dynamic covariate $N(t-)/t$ also included; dotted line, frailty model of Section 2.

young mother. Child age seems to be the most important with reduced incidence among older children.

Model fit within the Aalen class of models can be summarized by the empirical standard deviations of standardized martingale residuals, i.e. the difference between observed and expected counting processes, standardized by the model-based standard deviation. If the model is correctly specified then these should be close to one. Figure 2 shows these quantities for three different models: the Aalen model without heterogeneity, i.e. fixed covariates only; the Aalen model including a dynamic covariate (previous rate of episodes per unit time); and the frailty model fitted here. It is clear that the heterogeneity needs to be taken into account and there is little difference between the dynamic covariate and frailty approaches, although the latter may be slightly preferred.

Copyright © 2008 John Wiley & Sons, Ltd.

3. CLUSTERING

3.1. Identification of clusters

So far we have considered diarrhoea incidence to be independent among children. But because diarrhoea can be caused by bacteria or viruses that have been transmitted from person to person, in other words an infection, it is possible that there are clusters of cases. In this section, we cluster the children based on their living area and then we test whether there is a difference in diarrhoea patterns between these clusters. In Section 4, we will investigate whether any differences are time varying.

Since the data were obtained through daily home visits, we at first assumed that children assigned to the same data collector lived close to each other and clustered the data accordingly. We obtained the clusters by observing common missingness patterns for children with consecutive identification numbers, making the assumption that the missingness was caused by the data collector being unavailable. This suggested 21 clusters, with cluster sizes varying between 23 and 47. Subsequently, we received information on the childrens' home locations, which confirmed that the data were indeed comprised of samples from 21 different geographic districts within Salvador. Fortunately, the two clustering methods gave complete agreement. As an aside, we note that this suggests that our assumption of non-informative intermittent missingness is credible: the absence of an observation seems to be due to administrative reasons not associated with the health of the child.

3.2. Comparison of clusters

Before attempting modelling, we undertook some preliminary screening to determine whether there was an evidence of differences between the clusters. Figure 3 shows the fitted baseline cumulative functions for the 21 groups when each cluster was considered alone and covariates were not included. The baseline for a common fit is also shown. As a Monte Carlo test for between-cluster differences, we simulated data from the common baseline and clustered into 21 groups as for the original data. We then fitted a separate baseline intensity to each cluster in the simulated

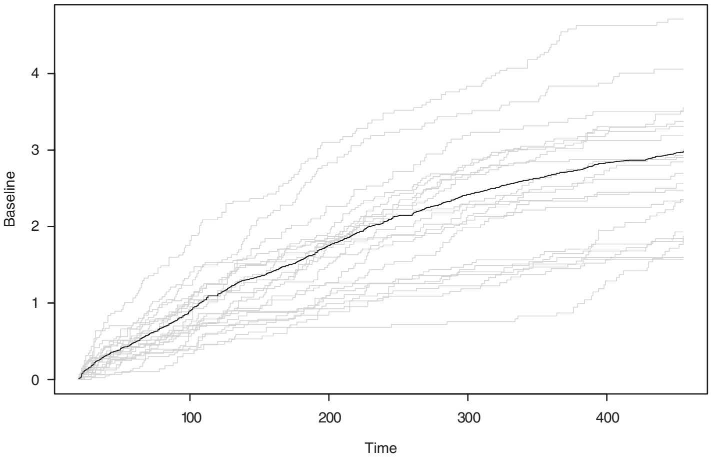

Figure 3. Cumulative baseline functions for the 21 clusters (grey lines) and for all children (black line).

Copyright © 2008 John Wiley & Sons, Ltd.

data and at each time point determined the between-cluster variance in these quantities. There was considerably more between-cluster variability in the diarrhoea data than can be explained by sampling error from a common model.

Some differences between clusters might be explained by demographic differences between districts, perhaps captured by known covariates. The procedure above does not take covariates into account and so far a more formal comparison with adjustment for covariates we used a log rank test for recurrent event data [11, 12]. The test is based on a weighted comparison at each event time between the observed and expected number of events in each cluster. We applied the test with weight $w(t)=1$ with and without covariates in the model. Having no covariates gave $\chi^2=143.568$ on 20 df, which has p-value <0.0001, and with covariates we obtained $\chi^2=118.752$ on 32 df, which also has p-value <0.0001. Thus, only part of the between-cluster differences is explained by differences in covariate patterns and we are confident that there is significant genuine association in the data.

3.3. *Correlated frailty model for recurrent event data*

Association in survival, or more general event-history applications, is often described by a shared frailty model under which intensities for all individuals in a cluster are affected by the same random effect and there is independence between clusters. Since we expect high individual heterogeneity in addition to cluster-level association we prefer to employ a correlated frailty model which allows partial association between-cluster members. This model was introduced by Petersen *et al.* [13] to model the clustered survival data. We extend the model here for clustered individuals with recurrent event data. Let $N(t)=((N_{i1}(t),\dots,N_{im_i}(t); i=1,\dots,n))$ be a multivariate counting process. In this set-up, n clusters with m_i subjects in the ith cluster are observed. The intensity process for subject j in cluster i is defined as (1) in which Z_i is replaced with $Z_i^{(j)}$ defined as follows:

$$\begin{aligned} Z_i^{(1)} &= Z_{i0}+Z_{i1} \\ \vdots \quad & \quad\quad \vdots \\ Z_i^{(m_i)} &= Z_{i0}+Z_{im_i} \end{aligned} \tag{4}$$

Here, $Z_{i0}, Z_{i1},\dots,Z_{im_i}$ are independent gamma distributed random variables with parameters $(v, 1/\xi), (v^*, 1/\xi),\dots,(v^*, 1/\xi)$, respectively. To make the intensity identifiable, we take $v+v^*=1/\xi$ so that $E(Z_i^{(j)}=1)$. Note that

$$\mathrm{Var}(Z_i^{(j)})=\mathrm{Var}(Z_{i0})+\mathrm{Var}(Z_{ij})=\xi$$

and the marginal distribution of $Z_i^{(j)}$ is gamma as for the frailty model of the previous section.

Independence between clusters is assumed but subjects within clusters are correlated through the common frailty Z_{i0}. The correlation is not complete because of the independent subject-specific terms Z_{ij}. A large value of v supports the clustering and a large value of v^* indicates high individual heterogeneity and less correlation within the clusters.

Estimation follows the algorithm of Section 2 with the exception that the frailty parameters (any two of v, v^*, and ξ) are estimated by maximizing for fixed $\beta_j(t)$ the cluster marginal likelihood defined in the Appendix.

Copyright © 2008 John Wiley & Sons, Ltd.

Table III. Diarrhoea data: tests for covariate effects under Aalen's additive model with correlated frailty.

β	Test U	SE(U)	U/SE	p-Value
Low socio-economic status	2.70	19.83	0.14	0.891
Other children $\leqslant$5 years	35.97	22.09	1.63	0.103
More than three people/bedroom	83.31	19.42	4.29	<0.001
Poor street quality	−24.68	19.27	−1.28	0.200
Contaminated water source	−25.35	18.43	−1.37	0.169
Contaminated water storage	21.32	19.23	1.11	0.267
Standing water	−0.55	17.93	−0.03	0.975
Open sewerage	91.01	16.08	5.66	<0.001
Rain affected accommodation	70.08	20.31	3.45	0.001
Mother <25 years	79.05	23.47	3.37	0.001
Gender	27.69	23.15	1.19	0.232
Starting age (months)	−7425.60	454.77	−16.33	<0.001

3.4. *Correlated frailty model applied to diarrhoea data*

When the correlated frailty model for recurrent event data was fitted to the diarrhoea data, we obtained $\hat{\xi}=0.84$ (bootstrap SE 0.139 from 40 samples), reassuringly close to the 0.81 obtained under the independent frailty model of the previous section. There is good evidence of clustering since the estimated variance of the shared component is 0.22, which is 27 per cent of total variance, leaving 0.62 (73 per cent) as the independent component variance. In the bootstrap samples, the proportion of frailty variance, which is explained by the shared component, varied from 23 to 39 per cent.

Test statistics for covariate effects are given in Table III. The same variables are significant as found in Table II, but evidence in all cases is stronger.

4. SPACE–TIME CLUSTERING

4.1. *Exploratory analysis*

Time-constant differences between clusters would indicate spatial or environmental effects not explained by covariates. If the difference between the clusters is due to local infections then we expect to see temporary periods of high risk within clusters. Figure 4 suggests that differences between clusters do indeed vary over time. The figure shows the average number of episodes per child-day for each month of the study and clusters are ordered according to the overall rate. Incidence of one episode per child per month, which would be a serious diarrhoea problem, corresponds to about 0.03 on this scale. We see that most clusters approach this level on occasions and no cluster is consistently very high or very low. Cluster 2, for instance, has low incidence for most of the study, but a relatively high rate appears in the closing months. Cluster 8 has the opposite pattern and there seems to be a similar high within-cluster variation for most clusters.

For a simple investigation, as to whether these differences can be explained by chance, a permutation study was used. First, we permuted all children and clustered them randomly into 21 clusters with sizes the same as the original data. Next we calculated again the average number of events per child-day in each new cluster every month. The between-cluster monthly variance

Copyright © 2008 John Wiley & Sons, Ltd.

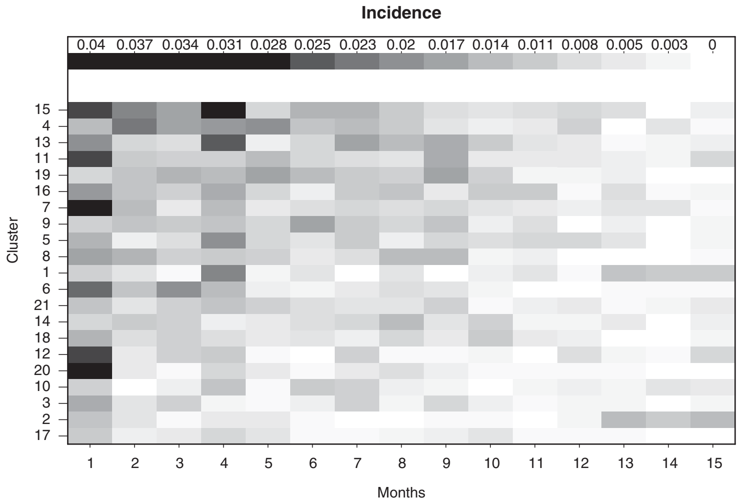

Figure 4. Cluster-specific monthly rates. Each cell represents the average number of events per child-day.

was then calculated and compared with the corresponding observed quantity. The procedure was repeated 1000 times and in all cases the observed variance was higher, each month, than the permutation variance. Hence further modelling is justified.

4.2. *Aalen additive model for space–time variation*

For a more formal approach to the identification of genuine time-varying cluster differences, we can fit the Aalen additive model with cluster-specific terms included through appropriate indicator variables. Cumulative coefficients, one per cluster, with standard errors are shown in Figure 5. Some clusters seem to have no real effect, some have fairly constant effects, and some have evidence of time variation.

Suppose the regression coefficient for cluster k is $\beta_k(t)$, $k=1,2,\ldots,21$. Then we will be interested in testing, for each cluster, the following hypotheses:

$$H_{01}\colon \beta_k(t)=0 \quad \text{for all } t \tag{5}$$

$$H_{02}\colon \beta_k(t)=\beta_k \quad \text{for all } t \tag{6}$$

First, we consider the null hypothesis $H_{01}\colon \beta_k(t)=0$ for all t, which is equivalent to test the hypothesis $H_{01}\colon B_k(t)=0$. If the hypothesis is true there is no difference between cluster k and the overall pattern. A test using standard machinery is based on a weighted average of the estimated coefficients [3–5]. Nine of these: clusters 1, 2, 3, 4, 10, 15, 17, 19, and 20 have highly significant differences from the overall pattern. In all the cases, comparison with Figure 5 shows that zero falls outside the pointwise confidence intervals for substantial regions of the time axis.

What is not clear is whether these differences are time constant or time varying, and hence the need to test the second of the hypotheses, H_{02}, of constant but non-zero effect. An innovative resampling test procedure developed by Scheike and Martinussen [11, 14] are summarized in the

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

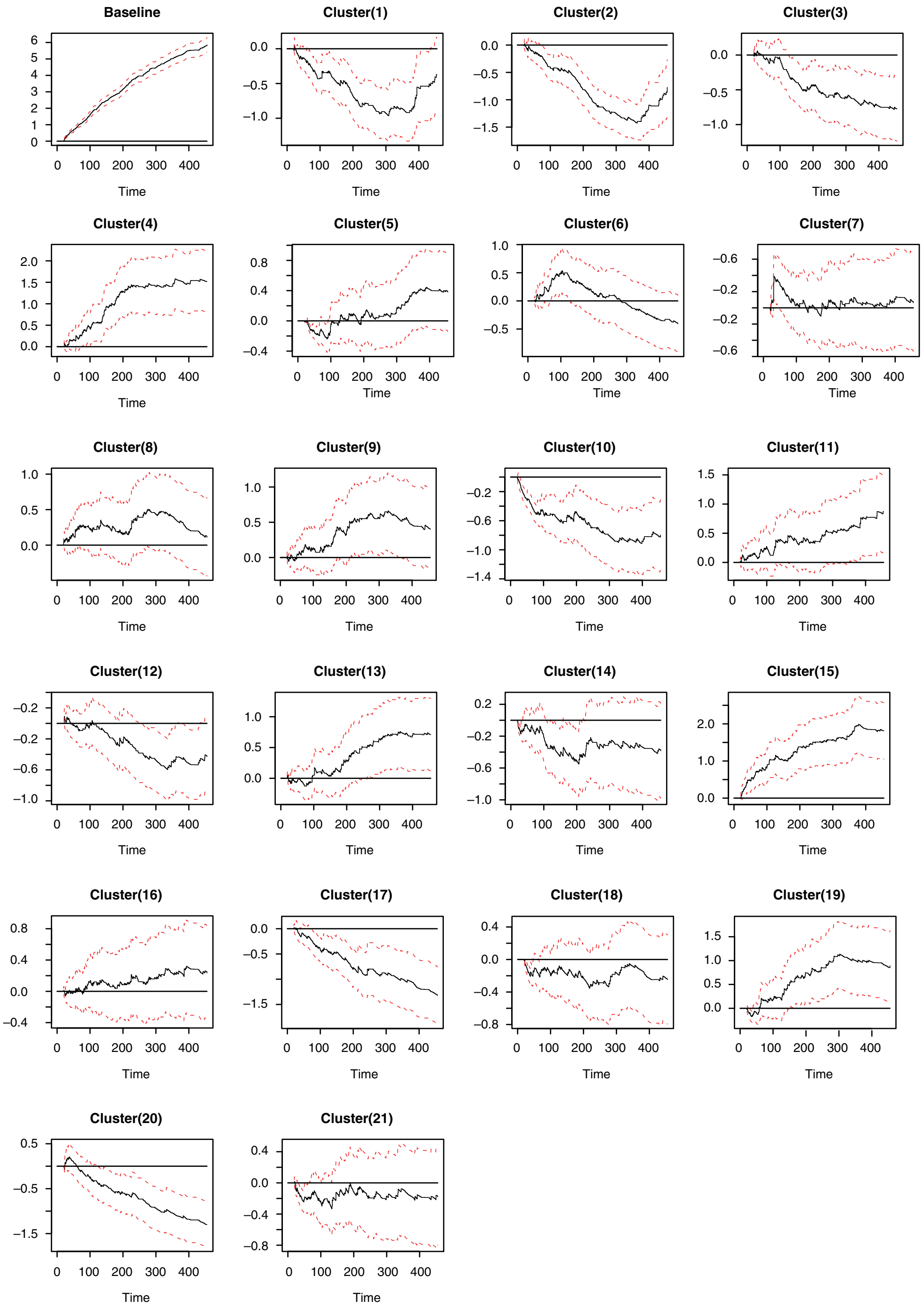

Figure 5. Cumulative coefficients $B_k(t)$ for the cluster effects, with ± two standard errors.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table IV. Estimated total, between- and within-cluster frailty variances over disjoint subintervals.

Time period	Variance		
days	Total	Between clusters	Within clusters
1–30	0.832	0.087	0.745
31–60	0.161	0.033	0.128
61–120	0.575	0.133	0.442
121–240	0.980	0.461	0.519
241–455	0.411	0.092	0.319
1–455	0.841	0.225	0.616

Appendix. The hypothesis of constant effect is rejected unambiguously only for clusters 1 and 2, though p-values are low also for clusters 4, 6, and 10. Clusters 1 and 2 have clear changes in slope of the cumulative coefficients (Figure 5) and from Figure 4 we see that they are the two clusters with relatively high incidence in the closing months of the study.

4.3. Correlated frailty model applied to subintervals

To close the analysis, we report results of fitting the correlated frailty model to disjoint subintervals of the axis. In Section 3, we assumed the frailty terms to be time constant and hence clusters were assumed to be consistently high or low throughout the study. With the results above in mind it seems more sensible to allow frailty to vary in time. Unfortunately, there seems as yet to be no available time-varying frailty model that can account for within-cluster correlation. Instead, we simply fit separate time-constant models to separate subintervals and to ensure that there are sufficient events for reliable estimation we choose interval lengths, which increase over time, as the number of episodes of diarrhoea correspondingly fall.

Estimated frailty parameters are given in Table IV and it is interesting to compare these with the rates illustrated in Figure 4. During the first month there is high variability between individuals but apparently little within-cluster association. In the second month, however, there is little variability as rates are low throughout. Between- and within-cluster variance pick up over the following months but declines again for the last six months or so.

5. DISCUSSION

We have described an event-history analysis of a longitudinal study into the incidence of diarrhoea among young children. We do not of course claim our analysis procedure to be the only approach to data of this type, since there are many alternatives available. But we do consider the method to have many advantages. The Aalen additive model and associated martingale basis for inference provide an easily computed and flexible basis for dealing with complex incidence data, with time-varying effects, dropout, and missing observations. The frailty extension reported here allows modelling of both heterogeneity not explained by covariates and clustering in both space and time, and is perhaps more intuitively reasonable than the dynamic covariate method

Copyright © 2008 John Wiley & Sons, Ltd.

[7, 8], where future incidence is assumed to depend on the previous history of events. This allows heterogeneity to develop over time, whereas the frailty approach assumes heterogeneity from the onset.

Further work to consider how a time-varying frailty structure can be incorporated would be useful. In Section 4, we simply divided the time interval into subintervals and assumed a separate frailty model for each. This is unrealistic since there is unlikely to be independence between clusters and also the partitioning is arbitrary and unlikely to be appropriate for all clusters. It would be preferable to introduce a smooth time-varying frailty $Z(t)$, either at the individual or at the cluster level, or both. How this can be achieved is not yet known, though a multivariate gamma distribution used in [15] for longitudinal count data may be a useful starting point.

Another topic for further work would be the conditional multiplier test developed by Scheike and Martinussen [11, 14] and described in the Appendix. The particular form of test we chose to use gives equal weight to all regions of the time axis, but it may be preferable to weight some regions more heavily than others, for instance, where there are more events. Similarly, it will be useful to compare power properties of different tests based on $\hat{B}(t)$. For instance, we have used the integrated squared difference between $\hat{B}(t)$ and $\hat{B}(\tau)t/\tau$, but [11] also describes a test based on the supremum of the absolute difference. How each test performs against particular alternatives is not known.

APPENDIX A

A.1. Marginal likelihood construction for correlated frailty model

First, we construct the likelihood for one cluster. For simplicity, take $\lambda_j(t)=\lambda_j(t|F_{t^-},x_j(t))$. For known frailties the likelihood for cluster i at time τ is

$$L_i(\beta,v,v^*,\xi)=\prod_{j=1}^{m_i}\left\{\prod_{t=1}^{\tau}[Z_i^{(j)}\lambda_j(t)]^{\mathrm{d}N_j(t)}\right\}\exp\left[-Z_i^{(j)}\int_0^{\tau}\lambda_j(u)\,\mathrm{d}u\right]$$

and the marginal likelihood can be proven to be

$$\begin{aligned}\mathrm{ML}_i(\beta,v,v^*,\xi)=&\left\{\prod_{j=1}^{m_i}\prod_{t=1}^{\tau}[\lambda_j(t)]^{\mathrm{d}N_j(t)}\right\}\left(\frac{1}{1+\xi\Lambda_\bullet(\tau)}\right)^{v}\prod_{j=1}^{m_i}\left(\frac{1}{1+\xi\Lambda_j(\tau)}\right)^{v^*}\\&\times\sum_{\mathbf{k}\in K(\tau)}\prod_{j=1}^{m_i}\left\{\frac{\Gamma(N_j(\tau)-k_j+v^*)}{\Gamma(v^*)}\left[\frac{\xi}{1+\xi\Lambda_j(\tau)}\right]^{N_j(\tau)-k_j}\right\}\\&\times\frac{\Gamma(k_\bullet+v)}{\Gamma(v)}\left[\frac{\xi}{1+\xi\Lambda_\bullet(\tau)}\right]^{k_\bullet}\end{aligned}$$

where $\Lambda_j(\tau)=\int_0^{\tau}\lambda(s)\,\mathrm{d}s$, $\Lambda_\bullet=\sum_{j=1}^{m_i}\Lambda_j(\tau)$, $N(\tau)=\sum_{t=1}^{\tau}\mathrm{d}N_j(t)$, $\mathbf{k}=(k_1,\ldots,k_{m_i})$, $K(u)=\{\mathbf{k}|k_j\in\{0,N_j(u)\},j=1,\ldots,m_i\}$, $k_\bullet=\sum_{j=1}^{m_i}k_j$.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

For the algorithm, we need the expected frailty terms given history to time t. The shared component is obtained as

$$E[Z_0(t|F_{t^-})]$$

$$=\frac{\xi}{1+\xi\Lambda_{\bullet}(t^-)}\frac{\sum_{\mathbf{k}\in K(t^-)}\prod_{j=1}^{m_i}\{\mathbf{C}_{k_j}^{N_j(t^-)}\Gamma(N_j(t^-)-k_j+v^*)\}\Gamma(k_{\bullet}+v+1)\left[\frac{\xi}{1+\xi\Lambda_{\bullet}(t^-)}\right]^{k_{\bullet}}}{\sum_{\mathbf{k}\in K(t^-)}\prod_{j=1}^{m_i}\{\mathbf{C}_{k_j}^{N_j(t^-)}\Gamma(N_j(t^-)-k_j+v^*)\}\Gamma(k_{\bullet}+v)\left[\frac{\xi}{1+\xi\Lambda_{\bullet}(t^-)}\right]^{k_{\bullet}}}$$

and the non-shared frailty is estimated as

$$E[Z_{j^*}(t|F_{t^-})]$$

$$=\frac{\xi}{1+\xi\Lambda_{j^*}(t^-)}\frac{\sum_{\mathbf{k}\in K(t^-)}\prod_{j=1}^{m_i}\{\mathbf{C}_{k_j}^{N_j(t^-)}k_j\Gamma(N_j(t^-)-k_j+v^*)\}(N_j(t^-)-k_{j^*}+v^*)\Gamma(k_{\bullet}+v)\left[\frac{\xi}{1+\xi\Lambda_{\bullet}(t^-)}\right]^{k_{\bullet}}}{\sum_{\mathbf{k}\in K(t^-)}\prod_{j=1}^{m_i}\{\mathbf{C}_{k_j}^{N_j(t^-)}\Gamma(N_j(t^-)-k_j+v^*)\}\Gamma(k_{\bullet}+v)\left[\frac{\xi}{1+\xi\Lambda_{\bullet}(t^-)}\right]^{k_{\bullet}}}$$

where $j^*\in\{j|j=1,\ldots,m_i\}$. The expected frailty is the sum of these terms.

A.2. Conditional multiplier test for time-constant effects

The following applies to any of the cumulative coefficients $B(t)$. We wish to test

$$H_{02}:\beta(t)=\beta \quad \text{for all } t$$

This hypothesis is equivalent to test $H_{02}:B(t)=\beta t$ for the cumulative regression coefficient. We will use $\hat{B}(\tau)/\tau$ as an estimate of the constant β under the null.

Because the process $n^{1/2}(\hat{B}(t)-\hat{B}(\tau)t/\tau)$ is not a martingale, standard methods based on $\hat{B}(t)$ and its variance cannot be used. Scheike [14] suggested instead a resampling approach, based on earlier work by Lin *et al.* [16], using so-called conditional multipliers. The idea is that, if $D_i, i=1,2,\ldots,n$, are independent with zero mean and $n^{-1/2}\sum_{i=1}^{n}D_i$ converges in distribution to some random variable D, and if Z_i are independent with zero mean and unit variance, then $\sum_{i=1}^{n}Z_iD_i$ also converges in distribution to D. Simulating Z_i thus gives us an approach to approximating the distribution of D.

First, we would need to write $\hat{B}(t)-B(t)$ as a sum of independent terms equivalent to $n^{-1/2}\sum_{i=1}^{n}D_i$. This is impossible. However, it can be shown [11, 14] that $\hat{B}(t)-B(t)$ *acts* in the required manner, which is sufficient. We summarize the argument here. Let $M(s)$ be a vector with elements $M_i(s)=N_i(s)-\Lambda_i(s)$. Then

$$\begin{aligned}\sqrt{n}\{\hat{B}(t)-B(t)\}&=\sqrt{n}\int_0^t(X^{\mathrm{T}}(s)X(s))^{-1}X(s)\,\mathrm{d}M(s)\\&=\sqrt{n}\int_0^t\left(\sum_i x_i(s)x_i^{\mathrm{T}}(s)\right)^{-1}\sum_i x_i(s)\,\mathrm{d}M_i(s)\\&=\frac{1}{\sqrt{n}}\sum_i\int_0^t\left(n^{-1}\sum_i x_i(s)x_i^{\mathrm{T}}(s)\right)^{-1}x_i(s)\,\mathrm{d}M_i(s)\\&=\frac{1}{\sqrt{n}}\sum_i\varepsilon_i(t)\end{aligned}$$

Copyright © 2008 John Wiley & Sons, Ltd.

say the inverse, with divisor n, converges at a faster rate than the remaining terms and can in effect be considered as deterministic so the sum acts like a sum of independent terms. Now let

$$\hat{\varepsilon}_i(t) = \int_0^t \left(n^{-1} \sum_i x_i(s) x_i^{\mathrm{T}}(s) \right)^{-1} x_i(s)\, \mathrm{d}\hat{M}_i(s)$$

and

$$\Delta(t) = n^{-1/2} \sum_{i=1}^{n} \hat{\varepsilon}_i(t) Z_i$$

where Z_i are independent and standard normally distributed. Then $\Delta(t)$ has the same limiting distribution as $n^{1/2}(\hat{B}(t) - B(t))$ [11].

Thus, a test statistic for $H_0\colon B(t) = \beta t$ can be constructed based on replications of $\Delta(t)$. The test statistic that is considered in our work is

$$T_1 = n \int_0^\tau \left(\hat{B}(t) - \hat{B}(\tau) \frac{t}{\tau} \right)^2 \mathrm{d}t$$

Under the null we have $B(\tau) = \beta\tau$ and $\hat{B}(t) - \hat{B}(\tau)t/\tau$ are equivalent to $\Delta(t) - \Delta(\tau)t/\tau$. Therefore, the function $n^{1/2}(\hat{B}(t) - \hat{B}(\tau)t/\tau)$ has the same asymptotic distribution of $\Delta(t) - \Delta(\tau)t/\tau$, and replications of T_1 based on the latter generate the appropriate sampling distribution.

ACKNOWLEDGEMENTS

The research of R. Fiaccone was supported in part by the National Counsel of Technological and Scientific Development—CNPq, Brazil.

REFERENCES

1. Barreto ML, Carmo E, Santos C, Ferreira L. 'Emergentes', 're-emergentes' e 'permanecentes': Tendancias recentes das doenas infecciosas e parasitarias no Brasil (in Portuguese). *Informe Epidemiolgico SUS* 1996; **3**:7–18.
2. Strina A, Cairncross S, Prado MS, Teles CA, Barreto ML. Childhood diarrhoea symptoms, management and duration: observations from a longitudinal community study. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2005; **99**:407–416.
3. Borgan Ø, Fiaccone RL, Henderson R, Barreto M. Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. *Scandinavian Journal of Statistics* 2007; **34**:53–69.
4. Aalen OO. A linear regression model for the analysis of life times. *Statistics in Medicine* 1989; **8**:907–925.
5. Aalen OO. Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine* 1993; **12**:1569–1588.
6. Pena EA, Strawderman RL, Hollander M. Nonparametric estimation with recurrent event data. *Journal of American Statistical Association* 2001; **99**:1299–1315.
7. Aalen OO, Fosen J, Wedon-Fekjær H, Borgan Ø, Husebye E. Dynamic analysis of multivariate failure time data. *Biometrics* 2004; **60**:764–773.
8. Fosen J, Borgan Ø, Weedon-Fekær H, Aalen OO. Dynamic analysis of recurrent event data using the additive model. *Biometrical Journal* 2006; **48**:381–398.
9. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**:439–454.
10. Lawless JF. Regression models for Poisson process data. *Journal of the American Statistical Association* 1987; **52**:808–815.
11. Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data.* Springer: Berlin, New York, 2006.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

12. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes.* Springer: Berlin, New York, 1993.
13. Peterson JH, Anderson PK, Gill RD. Variance components models for survival data. *Statistica Neerlandica* 1996; **50**:193–211.
14. Scheike TH. The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. *Lifetime Data Analysis* 2002; **8**:247–262.
15. Henderson R, Shimakura S. A serially correlated gamma frailty model for recurrent events. *Biometrika* 2003; **90**:355–366.
16. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**:73–81.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim