doi: 10.1111/j.1467-9469.2006.00525.x

© Board of the Foundation of the Scandinavian Journal of Statistics 2006. Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 34: 53–69, 2007

Dynamic Analysis of Recurrent Event Data with Missing Observations, with Application to Infant Diarrhoea in Brazil

ØRNULF BORGAN

Department of Mathematics, University of Oslo

ROSEMEIRE L. FIACCONE

Departamento de Estatistica, Universidade Federal da Bahia

ROBIN HENDERSON

School of Mathematics and Statistics, University of Newcastle upon Tyne

MAURICIO L. BARRETO

Instituto de Sa'ude Coletiva, Universidade Federal da Bahia

ABSTRACT. This paper examines and applies methods for modelling longitudinal binary data subject to both intermittent missingness and dropout. The paper is based around the analysis of data from a study into the health impact of a sanitation programme carried out in Salvador, Brazil. Our objective was to investigate risk factors associated with incidence and prevalence of diarrhoea in children aged up to 3 years old. In total, 926 children were followed up at home twice a week from October 2000 to January 2002 and for each child daily occurrence of diarrhoea was recorded. A challenging factor in analysing these data is the presence of between-subject heterogeneity not explained by known risk factors, combined with significant loss of observed data through either intermittent missingness (average of 78 days per child) or dropout (21% of children). We discuss modelling strategies and show the advantages of taking an event history approach with an additive discrete time regression model.

Key words: additive regression model, diarrhoea incidence and prevalence, discrete time martingales, dropout, longitudinal binary data, missing data

1. Introduction

Recurrent events are frequently of interest in longitudinal studies. Examples include seizures in epileptic patients (Albert, 1991) or successive tumours in cancer studies (Gail *et al.*, 1980). Approaches to the analysis of recurrent events include intensity-based counting process methods (Andersen *et al.*, 1993), the analysis of times to specific events (Wei *et al.*, 1989), times between events (Aalen & Husebye, 1991) and frailty modelling (Oakes, 1992; Yue & Chan, 1997). Miloslavsky *et al.* (2004) provide a recent overview of the methods used for recurrent event analyses.

In this work, we study additive dynamic regression models for discrete time recurrent event data in which the conditional mean based on the history is modelled as a function of possibly time-varying covariates. The paper is based on the analysis of data from an epidemiological study of the relationship between sanitation facilities and the occurrence of diarrhoea in children under 3 years old. We consider both days with diarrhoea and repeated episodes of diarrhoea as recurrent events and show how the armoury of additive regression modelling techniques developed for time continuous event history data (Aalen, 1989, 1993) may be applied to our longitudinal binary data to provide valuable inferences without computationally intensive procedures. Plots of the time-varying regression coefficients provide a useful graphical summary of the time dynamics of the covariate effects, and this makes the approach particularly important when individual experience of dynamic or changing conditions affects the occurrence of the recurrent events. For comparative purposes, we also consider a recently proposed but computationally intensive method for longitudinal binary data given by Albert (2000).

Details of the data to be considered are provided in the next section. In section 3 we describe a general modelling framework for discrete time recurrent event data subject to missingness, while the approach of Albert (2000) is briefly considered in section 4. Our additive regression model with dynamic covariates is introduced in section 5. Useful methods for statistical inference for the additive model are also reviewed and discussed in this section, while our analysis of the diarrhoea data using additive regression methods is given in section 6. The paper closes with discussion of open problems in section 7.

2. Blue Bay diarrhoea data

Poverty in many countries is associated with high risk of disease, in part related to poor sanitation and inadequate health education. Focusing on this topic, the Bahia state government (Brazil) has implemented an extensive sanitation programme since 1997 in the metropolitan area of Salvador. As part of the programme, the Institute of Public Health of the Federal University of Bahia developed several studies, together called *Blue Bay*, to evaluate the impact of the sanitation measures on the health of the population. In this paper, we will focus on the morbidity of diarrhoea in children up to 3 years of age.

Daily data are available from a household survey carried out through home visits over 455 days from October 2000 to January 2002. Study design and population have been described by Strina *et al.* (2005). One child aged under 3 years at entry was monitored from each household. In this work, we will concentrate on the 926 surveyed children who had at least 90 days of follow up, and we will investigate the *incidence* and *prevalence* of diarrhoea amongst these children through the period. Prevalence is the probability that a child has diarrhoea on a given day whereas incidence is the probability that a child starts a new episode of diarrhoea. An episode is a sequence of days with diarrhoea until there have been at least three consecutive clear days (diarrhoea free).

Figure 1 shows crude daily prevalence and incidence through the study period, computed as the proportions of children having diarrhoea, respectively, starting a new episode of

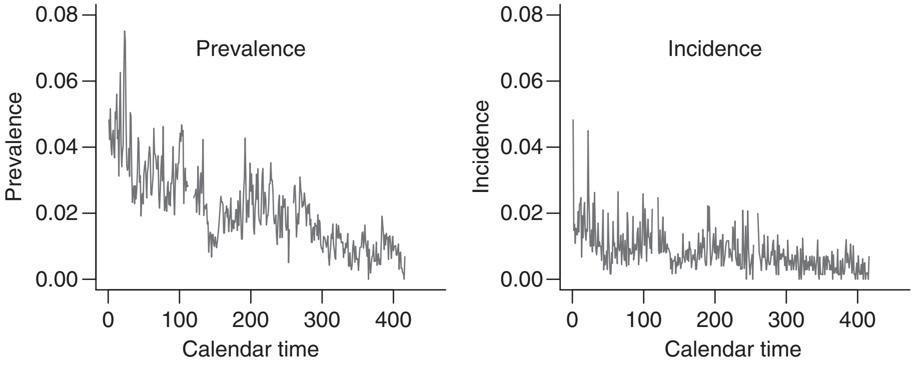

Fig. 1. Daily prevalence and incidence of diarrhoea after start of study.

Table 1. *Fixed covariate summary*

Description	Summary (%)
Male	47
Starting age (months)	
≤ 12	28
12–24	36
>24	36
Three or more people/bedroom	19
Poor street quality	57
Contaminated water storage	24
Contaminated water source	22
Standing water	32
Open sewerage	16
Rain-affected accommodation	29
Mother <25 yr	46
Low socio-economic status	61
Other children ≤ 5 yr	45

diarrhoea, on a given day. To begin with prevalence is around 5%, falling to about 1% 15 months later. Incidence by definition is lower, and is approximately 2% at the start of the study, 0.5% by the end. The fall in both plots may reflect improving health over the study period or may be an artefact due to the ageing of the cohort. Thus one of the challenges for the analysis is the need to disentangle calendar time and age effects, after allowance for other risk factors. Various social, demographic and economic characteristics were collected at the beginning of the study, many of which could influence outcome. Table 1 summarizes these covariates. In the analysis to come all these covariates except age are treated as binary, with the category shown in the table coded as 1. For age, three categories were considered, with the middle age group used as a reference. Daily data are also available on whether or not the child had vomit or fever.

A complication for the analysis is that all the children are not observed for the full study period. Figure 2 illustrates, by showing when children were and were not observed. The figure includes only every 10th child, as resolution becomes problematic with more dense

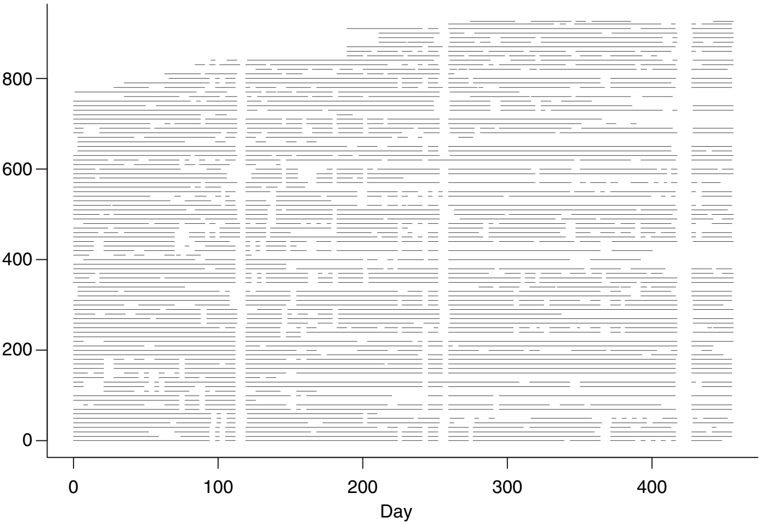

Fig. 2. Observation pattern for diarrhoea data. Horizontal lines indicate where data for each child are available, for every 10th child only.

data, but the pattern shown is entirely characteristic of the complete data. There are three types of missingness. First, some 16% of children were entered late into the study. Recruitment at the original start date of October 2000 was more problematic than anticipated and so a second recruitment phase took place from January to March 2001. This explains the mainly blank area in the top left of the plot. Secondly, about 21% of children dropped out of the study before the final completion date. Sometimes, this was for explained administrative reasons but some 15% were for unknown and potentially informative causes. The final cause of missing data was through intermittent missingness, whereby observation was interrupted for a period but later resumed. This was often because the data collector was not available, which is why there are many white vertical rectangles in Fig. 2. Data collectors were usually assigned blocks of children with contiguous identification numbers and if the data collector was not working through holiday or illness then data for the whole block was omitted. Often a small number of children have intermittent missing data but on four occasions there are almost no data at all, as seen by the vertical white bands running almost the full length of Fig. 2. Three are explained by vacation periods and the fourth happened during a strike by police.

Initial inspection of the data suggested that episodes of diarrhoea tend to be relatively short, but some children are more susceptible than others. This is confirmed by the lorelogram (Heagerty & Zeger, 1998) in Fig. 3, which gives the mean log odds ratios for 2×2 tables formed by the presence or absence of diarrhoea on days separated by given lags. Values bigger than zero imply positive association. There are two main features to this plot. At lag 1 the log odds ratio is very high, indicating not surprisingly that days with diarrhoea tend to follow each other. The lorelogram then decays very quickly for about 10 days, showing the episode effect. After that the mean is very stable at a level considerably above 0, which would be the value under independence. This long-term association occurs as a result of heterogeneity between children, essentially a frailty effect: some children have frequent episodes, some none or hardly any.

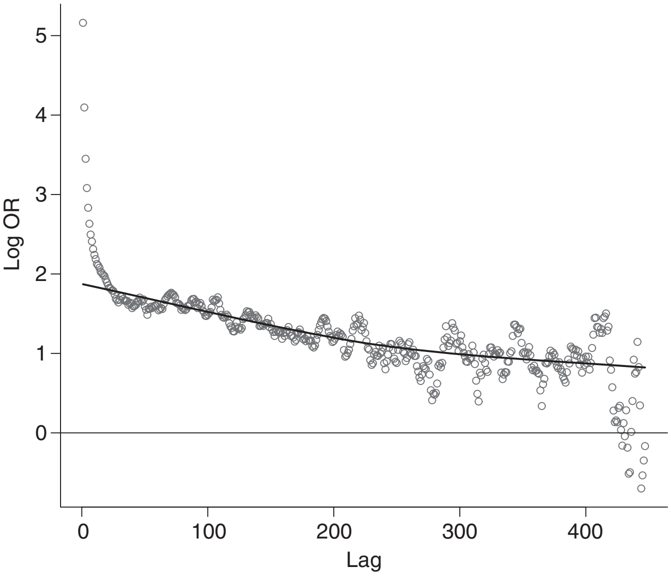

Fig. 3. Lorelogram for diarrhoea data: see text for explanation.

3. A modelling framework

We will consider the diarrhoea records for each child as longitudinal binary data, measured daily but subject to missingness, as discussed in the previous section. Using models in discrete time t, we will assume that $t \in \mathcal{T} = \{0, 1, \ldots, T\}$ for a given terminal time T. In our application, we will use days as the time unit and calendar time as the time scale, but we note that other time units and scales (such as years and age) may be more appropriate in other applications. In the following, we will consider two different types of models for the data: a transition model due to Albert (2000) and an additive model similar to the one proposed by Aalen (1980, 1989) for time-continuous event history data, and we will focus mainly on the latter. The two models will be described in sections 4 and 5. First, in this section, we introduce some notation and modelling assumptions common to both of them.

We start out by considering the hypothetical situation with no missing observations, which is the situation for which our basic model and parameters of interest are defined. For this situation our observations for the ith subject; $i = 1, \ldots, n$; form a binary process $\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iT}$, where $\tilde{Y}_{it} = 1$ if the individual experiences an event of interest at time t, $\tilde{Y}_{it} = 0$ otherwise. For completeness we let $\tilde{Y}_{i0} = 0$ so as to have $\tilde{Y}_{it}$ defined for all $t \in \mathcal{T}$. In our application the event of interest will be the onset of an episode of diarrhoea (when incidence is studied) or that the child suffers from diarrhoea (when prevalence is studied).

In addition, for each individual we at each time t have a p-dimensional vector of covariates $\mathbf{x}_{it} = (x_{i1t}, \ldots, x_{ipt})^{\mathrm{T}}$. These may be fixed or vary with time. For the transition model of section 4, all time-dependent covariates are assumed to be *external*, while also *dynamic* time-dependent covariates are allowed for the additive model of section 5. A time-dependent covariate is external if its complete path x_{ijt}; $t \in \mathcal{T}$; is given at the outset of the study or if its path is given by a stochastic process whose development over time is *not* influenced by the $\tilde{Y}_{it}$ (Kalbfleisch & Prentice, 2002, section 6.3). In both cases we may, for the purpose of statistical modelling, assume that the complete covariate paths are given at $t = 0$. In contrast, a dynamic time-dependent covariate may depend in an arbitrary way on 'the past', i.e. x_{ijt} may be a function of $\tilde{Y}_{is}$, for $s = 0, 1, \ldots, t-1$, as well as of the fixed and external time-varying covariates (Aalen *et al.*, 2004). Specific examples of dynamic covariates are given in section 6.1.

We denote by $\mathcal{H}_{i0}$ the σ-algebra generated by the fixed and external time-varying covariates for the ith subject, and let $\mathcal{H}_{it} = \mathcal{H}_{i0} \vee \sigma\{\tilde{Y}_{i1}, \tilde{Y}_{i2}, \ldots, \tilde{Y}_{it}\}$. Note that $\mathcal{H}_{it}$ may be interpreted as the information on the ith subject that would have been available by time t had there been no missing observations, assuming the complete path of external time-varying covariates to be known at the outset of the study. Then, conditional on $\mathcal{H}_{i0}$, the joint distribution of $\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iT}$ may be given by the *conditional* probabilities

$$\alpha_{it} = P(\tilde{Y}_{it} = 1 \mid \mathcal{H}_{i,\, t-1}). \tag{1}$$

A main aim for our analysis of the longitudinal binary data is to study how these conditional probabilities vary over time and how they depend on covariates. Note that this differs from the common approach in longitudinal data analysis, where focus is on estimating the *marginal* probabilities $\mu_{it} = P(\tilde{Y}_{it} = 1 \mid \mathcal{H}_{i0})$; e.g. Hogan *et al.* (2004).

The study of the α_{it} is complicated by missing observations. In order to handle the missingness, we introduce the categorical 'missingness process' $Z_{i1}, \ldots, Z_{iT}$, where Z_{it} indicates whether the outcome $\tilde{Y}_{it}$ for subject i is observed, lost due to intermittent missingness or lost due to dropout:

$$Z_{it} = \begin{cases} 0, & \text{observed} \\ 1, & \text{intermittent missing} \\ 2, & \text{dropout.} \end{cases}$$

In order to have Z_{it} defined for all $t \in \mathcal{T}$, we let $Z_{i0}=0$. We assume that Z_{it} is observed at time t, so that one knows 'today' whether a missing observation is intermittent or due to drop out. This assumption is fulfilled when intermittent missingness is caused by a data collector not being available or when drop-out is caused by a family moving out of the study area (provided the information is recorded in the data), but it may be more problematic otherwise.

The introduction of the missingness process will (usually) bring in some extra random variation. Therefore, we now have to work with the larger filtration $(\mathcal{G}_{it})$ given by the σ-algebras

$$\mathcal{G}_{it}=\mathcal{G}_{i0} \vee \sigma\{Z_{i1}, \tilde{Y}_{i1}, Z_{i2}, \tilde{Y}_{i2}, \dots, Z_{it}, \tilde{Y}_{it}\}.$$

Here, $\mathcal{G}_{i0}$ is generated both by the fixed and external time-varying covariates for subject i (i.e. $\mathcal{H}_{i0}$) and by those aspects of the missingness process for the subject that are external to its event process (as when an investigator misses a home visit for reasons that have nothing to do with the health condition of a child). This may have the consequence that the conditional distribution of $\tilde{Y}_{it}$ may change. It is, however, a basic assumption for our analysis that this is not the case, so that the missingness process satisfies

$$P(\tilde{Y}_{it}=1 \mid \mathcal{G}_{i,t-1})=P(\tilde{Y}_{it}=1 \mid \mathcal{H}_{i,t-1}) \tag{2}$$

for all $t \in \mathcal{T}$. Condition (2) is similar to one of the two conditions needed for independent censoring in event history analysis (Andersen *et al.*, 1993, sections III.2.2 and III.4). For the other condition, see section 5.1 below.

Under (2), conditional on fixed and external time-varying covariates as well as external aspects of the missingness process (i.e. on $\mathcal{G}_{i0}$), the joint distribution of $\tilde{Y}_{i1}, \dots, \tilde{Y}_{iT}, Z_{i1}, \dots, Z_{iT}$ may be given by the α_{it} and the conditional missingness probabilities

$$P(Z_{it}=m \mid \mathcal{G}_{i,t-1}, \tilde{Y}_{it}=y); \quad m=0,1,2; \ y=0,1. \tag{3}$$

Individuals may share values of fixed and external time-varying covariates and external aspects of the missingness processes. Thus it is not reasonable to assume independence of the n individuals. We will, however, assume that the vectors $(\tilde{Y}_{i1}, \dots, \tilde{Y}_{iT}, Z_{i1}, \dots, Z_{iT})$; $i=1, \dots, n$; are independent, conditional on all the $\mathcal{G}_{i0}$. Then the (conditional) model for all the n individuals may be specified by the α_{it} and the conditional missingness probabilities (3). The conditional independence assumption disregards all dependence between individuals that are not captured by observables, and this makes the assumption debatable for a contagious disease like diarrhoea; cf. the discussion in section 7.

4. A transition model

We now consider more closely the transition model proposed by Albert (2000). As discussed in the previous section, we for this model have to assume that all time-dependent covariates are external. Then, conditional on fixed and external time-varying covariates as well as external aspects of the missingness process (i.e. on $\mathcal{G}_{i0}$), Albert assumed Markov models for the event and missingness processes. For the event processes he assumed the logistic model

$$\operatorname{logit}(\alpha_{it})=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{it}+\theta \tilde{Y}_{i,t-1}.$$

Note that α_{it} depends on 'the past' $\mathcal{G}_{i,t-1}$ only via the covariates and $\tilde{Y}_{i,t-1}$, making the model for the longitudinal binary data Markovian. Higher-order Markov-dependence models could be assumed, at the cost of a dramatic increase in the computational burden.

To model the missingness probabilities (3), Albert assumed dependence on ‘the past’ only through the value of $Z_{i,\,t-1}$ and adopted the multinomial logit model:

$$P(Z_{it}=m\,|\,Z_{i,\,t-1}=l,\,\tilde{Y}_{it}=y)=\frac{\exp(\gamma_{lm}^{\mathrm{T}}\mathbf{x}_{it}+\eta_{lm}y)}{\sum_{k=0}^{2}\exp(\gamma_{lk}^{\mathrm{T}}\mathbf{x}_{it}+\eta_{lk}y)};\quad l,m=0,1,2;y=0,1.$$

Note that the dependence between the event process $\tilde{Y}_{it}$ and the missingness process Z_{it} arises through the inclusion of the value of $\tilde{Y}_{it}$ in the missingness model.

Albert proposed an expectation maximization (EM) algorithm for estimation and gave a recursive estimation procedure for calculation of the conditional probability distribution of missing $\tilde{Y}_{it}$, given the observed data. In our case, with occasional reasonably long sequences of intermittently missing data, we found the recursive procedure to be unreliable thanks to accumulating numerical inaccuracies. Instead we found a Monte Carlo EM procedure to work well, tested by simulations, using Gibbs sampling to fill in missing values and averaging over iterations to estimate the required expectations. As Gibbs is used, we only need to generate any missing $\tilde{Y}_{it}$ given its immediate neighbours, which are generated sequentially if also missing. Standard errors (SE) were estimated by bootstrap with 100 resamples.

Table 2 shows the estimates and SE for the events model and for the three types of transition between Z values. We took events to be days with diarrhoea and so the results relate to prevalence. For the events model, young children are more prone to diarrhoea than older, as expected, and the risk of diarrhoea is higher in houses which are affected by rain or near open sewers. Children of younger mothers, with less experience, tend to have more diarrhoea and there is also increased risk in more crowded accommodation. To investigate calendar time effect, we partitioned the study period into three intervals, namely 0–150 days, 151–300 days, and over 300 days, with the first group as reference and dummy variables for the others. There was strong evidence of decrease in frequency as time proceeded, as anticipated. Finally for this analysis, we found the previous binary response to be highly predictive, again as expected.

Turning briefly to the missing data models, a variety of covariates appeared to be important in affecting transitions. These are not discussed in detail but we note from the last row of the table that the parameter which characterizes the dependence between the outcome $\tilde{Y}$

Table 2. *Estimates and standard errors (SE) for transition model*

		Missing data model		
Covariates	Events model	$m=1$ and $l=0$	$m=2$ and $l=0$	$m=0$ and $l=1$
Male	0.08 (0.05)	0.04 (0.03)	0.02 (0.19)	0.02 (0.03)
Age				
≤ 12 months	0.24 (0.07)	−0.15 (0.05)	0.79 (0.36)	−0.02 (0.05)
>24 months	−0.56 (0.13)	0.03 (0.03)	0.02 (0.26)	−0.01 (0.03)
≥ 3 people/bedroom	0.26 (0.08)	−0.02 (0.04)	0.09 (0.25)	−0.01 (0.04)
Poor street quality	−0.08 (0.05)	0.04 (0.03)	0.002 (0.19)	0.03 (0.03)
Contaminated water storage	−0.05 (0.06)	0.17 (0.03)	0.08 (0.20)	0.05 (0.03)
Contaminated water source	0.11 (0.06)	−0.02 (0.03)	−0.04 (0.23)	0.02 (0.03)
Standing water	0.01 (0.07)	0.01 (0.04)	−0.41 (0.30)	−0.002 (0.04)
Open sewerage	0.37 (0.10)	0.10 (0.05)	0.36 (0.26)	−0.02 (0.04)
Rain-affected accommodation	0.14 (0.06)	−0.05 (0.04)	−0.12 (0.20)	0.07 (0.03)
Mother <25 yr	0.17 (0.06)	−0.04 (0.03)	0.30 (0.18)	0.002 (0.03)
Low socio-economic status	−0.002 (0.05)	−0.27 (0.04)	−0.48 (0.17)	−0.01 (0.03)
Other children $\leq$ 5 yr	−0.02 (0.04)	−0.07 (0.03)	−0.17 (0.16)	0.04 (0.03)
Period				
150–300 d	−0.14 (0.04)	0.02 (0.03)	0.88 (0.24)	0.18 (0.03)
>300 d	−0.60 (0.08)	−0.28 (0.04)	1.37 (0.49)	−0.13 (0.03)
Diarrhoea previous day (θ)	4.92 (0.39)			
Diarrhoea current day (η)		−0.18 (0.59)	−0.04 (3.40)	0.02 (0.63)

and the missing data mechanism was not found to be significant for any transition, suggesting that intermittent missingness and dropout are both non-informative. Further details of this analysis are omitted.

5. An additive model

We now turn to the additive model for longitudinal binary data. As discussed in section 3, we for this model allow the time-dependent covariates for an individual to be dynamic, i.e. to depend on the past of its event process. The additive model is given by

$$\alpha_{it} = \beta_{0t} + \beta_{1t} x_{i1t} + \cdots + \beta_{ipt} x_{ipt}. \tag{4}$$

Note that in (4) the regression parameters β_{jt} are allowed to depend on time, giving the model a non-parametric flavour. In fact, our additive model is a discrete time version of Aalen's (1980, 1989) non-parametric additive hazards model for continuous time event history data. As we will see below, most of the methods of statistical inference for Aalen's model apply with only minor modifications to our situation with time discrete longitudinal binary data.

5.1. Modelling the observable data

In section 3, we introduced the filtrations $(\mathcal{H}_{it})$ and $(\mathcal{G}_{it})$ corresponding, respectively, to the situation with no missing observations and the situation where both the event process and the missingness process for subject i are observed. None of these filtrations describe the information actually available to the researcher. We will use martingale methods to study statistical methods for the additive model (4). Then we need to consider the filtration $(\mathcal{F}_{it})$ corresponding to the data actually available to the researcher on the ith subject; $i = 1, \ldots, n$.

To this end we introduce the 'at-risk' process $R_{it} = I\{Z_{it} = 0\}$ taking the value 1 if individual i is observed at time t and the value 0 otherwise, and the process $Y_{it} = R_{it}\tilde{Y}_{it}$, registering the observed events for the individual. The R_{it} process corresponds to a filtering process for event history models, while Y_{it} corresponds to (the increments of) a filtered counting process (Andersen *et al.*, 1993, section III.4). For the time-continuous case, it is common to assume the filtering process to be predictable. In a similar manner it is useful to formulate the discrete time problem in such a way that R_{it} becomes a predictable process relative to the observed filtration $(\mathcal{F}_{it})$. This is achieved by letting

$$\mathcal{F}_{it} = \mathcal{G}_{i0} \vee \sigma\{Z_{i1}, Y_{i1}, Z_{i2}, Y_{i2}, \ldots, Z_{it}, Y_{it}, Z_{i,t+1}\}, \tag{5}$$

so that $\mathcal{F}_{it}$ contains information of the missingness process 'one day ahead'. Note that in (5) there is an implicit assumption that all fixed and external time-varying covariates are observable, so that the information in $\mathcal{G}_{i0}$ is available to the researcher.

Unlike the case for the transition model of section 4, we assume for the additive model that $\tilde{Y}_{it}$ and Z_{it} are conditionally independent given $\mathcal{G}_{i,t-1}$; essentially this amounts to assuming sequential missingness at random (e.g. Hogan *et al.*, 2004). Note that this assumption is trivially fulfilled when missingness is external to the event process, as seems to be the case for intermittent missingness for the diarrhoea data (Fig. 2). Also the dropout in the diarrhoea data seem to be missing sequentially at random by the transitional analysis (Table 2).

We further assume that the dynamic time-dependent covariates in (4) depend only on the parts of the information in $\mathcal{G}_{i,t-1}$ that are contained also in $\mathcal{F}_{i,t-1}$; an assumption that is needed if the covariates are to be used for statistical modelling. Then it follows by (1) and (2) that

$$\begin{aligned}\lambda_{it} &= P(Y_{it}=1\,|\,\mathcal{F}_{i,\,t-1}) = \mathrm{E}\{\mathrm{E}(R_{it}\tilde{Y}_{it}\,|\,\mathcal{G}_{i,\,t-1}, Z_{it})\,|\,\mathcal{F}_{i,\,t-1}\}\\ &= R_{it}\mathrm{E}\{P(\tilde{Y}_{it}=1\,|\,\mathcal{G}_{i,\,t-1}, Z_{it})\,|\,\mathcal{F}_{i,\,t-1}\} = R_{it}\mathrm{E}\{\alpha_{it}\,|\,\mathcal{F}_{i,\,t-1}\}\\ &= \alpha_{it}R_{it}. \end{aligned} \tag{6}$$

Note that the conditional probabilities (6) for the actually observed binary data Y_{it} coincide with the conditional probabilities (1) for the fully observed binary data $\tilde{Y}_{it}$ whenever the former are observable. This corresponds to independent censoring for time-continuous event history models (Andersen *et al.*, 1993, sections III.2.2 and III.4), and shows that the sequential missingness at random assumption corresponds to the predictability of the missingness process, which is the second assumption needed for independent censoring for time-continuous event history models.

For ease of exposition we have assumed that the filtrations $(\mathcal{F}_{it})$ corresponding to the data actually available to the researcher take the form (5). Sometimes one may want to work with larger filtrations, that are also generated by other processes observed in parallel with the longitudinal binary data. For example, in the diarrhoea study, vomit and fever were also recorded for each child at the home visits. As long as prediction is not a concern, such an extension of the filtrations causes no problems for the statistical methods for the additive model, and we will also use the notation $(\mathcal{F}_{it})$ when the filtrations are enlarged.

Another comment is also in order concerning the filtrations $(\mathcal{F}_{it})$; $i=1,\dots,n$. These generate a common filtration $(\mathcal{F}_t)$ for all the individuals, and formally it would have been more correct to define conditional probabilities and expectations with respect to this common filtration. However, due to the conditional independence assumption (given the σ-algebra generated by the $\mathcal{G}_{i0}$) we have chosen not explicitly to do so.

5.2. Inference for the additive model

We now turn to estimation in the additive model (4). To this end we use the Doob decomposition for discrete time martingales to decompose Y_{it} into the sum of a systematic part $\lambda_{it}=E(Y_{it}\,|\,\mathcal{F}_{i,\,t-1})=\alpha_{it}R_{it}$ and a random error $\epsilon_{it}=Y_{it}-\lambda_{it}$. Here the ϵ_{it} are martingale differences, i.e. the process $M_{it}=\sum_{s=0}^{t}\epsilon_{is}$ is a martingale. Therefore, by (4), we may write

$$Y_{it}=\beta_{0t}R_{it}+\beta_{1t}x_{i1t}R_{it}+\cdots+\beta_{pt}x_{ipt}R_{it}+\epsilon_{it}, \tag{7}$$

which, for each t, has the form of a linear regression model with uncorrelated errors. We may therefore estimate the β_{jt} by regressing the observations Y_{it} on the covariates $x_{ijt}R_{it}$ using ordinary or weighted least squares. Although the estimates at each time point will be subject to fairly large sampling errors, one may obtain stable and informative estimates of the cumulative regression coefficients $B_{jt}=\sum_{s=0}^{t}\beta_{js}$ by accumulating the estimates of the β_{js} over time.

To describe in more detail how the estimation is carried out, it is convenient to introduce vector and matrix notation. For each $t\in\mathcal{T}$ we let $\mathbf{Y}_t=(Y_{1t},\dots,Y_{nt})^{\mathrm{T}}$ be the vector of observations, $\boldsymbol{\beta}_t=(\beta_{0t},\beta_{1t},\dots,\beta_{pt})^{\mathrm{T}}$ the vector of regression coefficients, $\mathbf{X}_t$ the ‘design matrix’ with rows $\mathbf{x}_{it}^{\mathrm{T}}R_{it}=(1,x_{i1t},\dots,x_{ipt})R_{it}$, and $\mathbf{W}_t=\mathrm{diag}\{w_{it}\}$ a diagonal matrix of predictable weights. Then, provided $\mathbf{X}_t$ has full rank, the weighted least squares estimate for $\boldsymbol{\beta}_t$ becomes $\hat{\boldsymbol{\beta}}_t=(\mathbf{X}_t^{\mathrm{T}}\mathbf{W}_t\mathbf{X}_t)^{-1}\mathbf{X}_t^{\mathrm{T}}\mathbf{W}_t\mathbf{Y}_t$. Let J_t be an indicator process taking the value 1 if $\mathbf{X}_t$ has full rank, and the value 0 otherwise. By accumulating the least squares estimates for all times when estimation is meaningful, we obtain the estimate

$$\hat{\mathbf{B}}_t=\sum_{s=0}^{t}J_s\hat{\boldsymbol{\beta}}_s=\sum_{s=0}^{t}J_s\left(\mathbf{X}_s^{\mathrm{T}}\mathbf{W}_s\mathbf{X}_s\right)^{-1}\mathbf{X}_s^{\mathrm{T}}\mathbf{W}_s\mathbf{Y}_s \tag{8}$$

for the vector of cumulative regression functions $\mathbf{B}_t=(B_{0t},B_{1t},\dots,B_{pt})^{\mathrm{T}}$.

To study the properties of this estimator, we introduce $\mathbf{B}_t^* = \sum_{s=0}^t J_s \boldsymbol{\beta}_s$, which is close to $\mathbf{B}_t$ when there is only a small probability that $\mathbf{X}_s$ does not have full rank for all $s \leq t$, and let $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})^{\mathrm{T}}$ be the vector of random errors in (7). Then $\mathbf{Y}_s = \mathbf{X}_s \boldsymbol{\beta}_s + \boldsymbol{\epsilon}_s$, and inserting this in (8) we obtain

$$\hat{\mathbf{B}}_t - \mathbf{B}_t^* = \sum_{s=0}^t J_s (\mathbf{X}_s^{\mathrm{T}} \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}_s^{\mathrm{T}} \mathbf{W}_s \boldsymbol{\epsilon}_s.$$

Thus $\hat{\mathbf{B}}_t - \mathbf{B}_t^*$ is a martingale transformation (the discrete time analogue of a stochastic integral), and hence a mean zero (vector-valued) martingale. In particular, $E\hat{\mathbf{B}}_t = E\mathbf{B}_t^*$ for all $t \in \mathcal{T}$, so (8) is almost an unbiased estimator. By a slight modification of the argument reviewed in Andersen *et al.* (1993, section VII.4.1) for Aalen's additive model for time-continuous event history data, one may show that the covariance matrix of $\hat{\mathbf{B}}_t$ may be estimated by

$$\widehat{\text{cov}}(\hat{\mathbf{B}}_t) = \sum_{s=0}^t J_s (\mathbf{X}_s^{\mathrm{T}} \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}_s^{\mathrm{T}} \mathbf{W}_s \hat{\boldsymbol{\Sigma}}_s \mathbf{W}_s \mathbf{X}_s (\mathbf{X}_s^{\mathrm{T}} \mathbf{W}_s \mathbf{X}_s)^{-1}, \tag{9}$$

where $\hat{\boldsymbol{\Sigma}}_s = \text{diag}\{\hat{\lambda}_{is}(1 - \hat{\lambda}_{is})\}$ is the $n \times n$ diagonal matrix with ith diagonal element equal to $\hat{\lambda}_{is}(1 - \hat{\lambda}_{is})$ with

$$\hat{\lambda}_{is} = \mathbf{x}_{it}^{\mathrm{T}} R_{it} \hat{\boldsymbol{\beta}}_t = \{\hat{\beta}_{0s} + \hat{\beta}_{1s} x_{i1s} + \cdots + \hat{\beta}_{ps} x_{ips}\} R_{is} \tag{10}$$

a model-based estimate of λ_{it}; cf. (4) and (6). Moreover, by the martingale central limit theorem, (8) is approximately multivariate normally distributed in large samples. Also a test for the hypothesis that a covariate has no effect, can be derived in a similar manner as for Aalen's additive model for time-continuous event history data, and we omit the details.

The estimator (9) of the covariance matrix of $\hat{\mathbf{B}}_t$ is valid when our model for $\lambda_{it} = E(Y_{it} \mid \mathcal{F}_{i,t-1})$ adequately describes its dependence on 'the past' $\mathcal{F}_{i,t-1}$. In particular this requires that the dynamic covariates used in (4) capture (most of) this dependence. Alternatively, we may resort to a marginal model, just assuming

$$E(Y_{it} \mid R_{it}, \mathbf{x}_{it}) = \mathbf{x}_{it}^{\mathrm{T}} R_{it} \boldsymbol{\beta}_t = \beta_{0t} R_{it} + \beta_{1t} x_{i1t} R_{it} + \cdots + \beta_{pt} x_{ipt} R_{it}.$$

Then, if the individuals are independent, we may copy the argument of Scheike (2002) to get the estimator

$$\widetilde{\text{cov}}(\hat{\mathbf{B}}_t) = \sum_{i=1}^n \mathbf{Q}_{it}^{\otimes 2} \tag{11}$$

for the covariance matrix of (8). Here, for a vector $\mathbf{a}$, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ and

$$\mathbf{Q}_{it} = \sum_{s=0}^t J_s (\mathbf{X}_s^{\mathrm{T}} \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{x}_{is} w_{is} (Y_{is} - \hat{\lambda}_{is}).$$

5.3. Martingale residual processes

One important tool to assess the fit of an additive model, is inspection of the martingale residual processes. These were introduced by Aalen (1993) in the context of his additive model for time-continuous event history data, and their use for recurrent event data was illustrated by Aalen *et al.* (2004). We will here briefly consider the martingale residual processes for longitudinal binary data in discrete time.

To this end we, for each individual i, introduce the process $N_{it} = \sum_{s=0}^t Y_{is}$ counting the number of observed events for the individual up to and including time t, and the process

$\Lambda_{it} = \sum_{s=0}^{t} \lambda_{is}$. Then $M_{it} = N_{it} - \Lambda_{it}$ is a martingale. The idea is now to replace Λ_{it} by its estimate $\hat{\Lambda}_{it} = \sum_{s=0}^{t} \hat{\lambda}_{is}$ under the model [cf. (10)] to obtain the martingale residual process

$$\hat{M}_{it} = N_{it} - \hat{\Lambda}_{it}. \tag{12}$$

In a similar manner as in Aalen (1993), we may show that $\hat{\mathbf{M}}_t = (\hat{M}_{1t}, \ldots, \hat{M}_{nt})^{\mathrm{T}}$ is a mean zero vector-valued martingale when the model is correctly specified, and that its covariance matrix may be estimated by

$$\widehat{\mathrm{cov}}(\hat{\mathbf{M}}_t) = \sum_{s=0}^{t} J_s(\mathbf{I} - \mathbf{H}_s)\hat{\boldsymbol{\Sigma}}_s(\mathbf{I} - \mathbf{H}_s)^{\mathrm{T}}. \tag{13}$$

Here $\mathbf{H}_s = \mathbf{X}_s(\mathbf{X}_s^{\mathrm{T}}\mathbf{W}_s\mathbf{X}_s)^{-1}\mathbf{X}_s^{\mathrm{T}}\mathbf{W}_s$ is the 'hat matrix', and $\hat{\boldsymbol{\Sigma}}_s$ is given just above (10).

We may now derive standardized martingale residual processes by dividing each process (12) by the square root of the corresponding diagonal element of (13). If the model is correctly specified, the standardized martingale residual processes should have mean 0 and variance 1. Following Fosen *et al.* (2006) we will check the fit of a model in sections 6.3 and 6.4 by plotting the empirical standard deviation of the standardized residual processes as a function of time. If a model fits reasonably well, the empirical standard deviation should be about 1, while larger values indicate a poor fitting model.

6. Analysis of the Blue Bay data

We now present our analysis of the Blue Bay diarrhoea data using the methods for additive regression described in the previous section. We start out with a discussion of the fixed and dynamic covariates used in the analysis, and then give the results for the analysis of dropouts, incidence and prevalence. For the dropout analysis we used an unweighted analysis. For incidence and prevalence we weighted by the inverse probability of not dropping out at the next time point. These weights were obtained from the dropout analysis.

6.1. Fixed and dynamic covariates

Table 1 summarizes the fixed covariates used in the analyses. In all except one case we used a binary coding, with the category coded as 1 shown in the table. The exception is age, where we used either the exact value or the three-group categorization given in the table, with 12–24 months as reference category. In both cases we incremented age as time proceeded, so the interpretation is as the age effect on any given day, not age at the beginning of the study. In the following we report only the analyses with categorized age: those with exact age are broadly similar.

We defined dynamic covariates as the historical subject-specific rate of episodes, days with diarrhoea, days with fever and days with vomit. More precisely, in each of these four cases we defined a dynamic covariate x_{ijt} for individual i as

$$x_{ijt} = \frac{\sum_{s=0}^{t-1} w_s R_{is} \tilde{Y}_{is}}{\sum_{s=0}^{t-1} w_s R_{is}} = \frac{\sum_{s=0}^{t-1} w_s Y_{is}}{\sum_{s=0}^{t-1} w_s R_{is}},$$

where $\tilde{Y}_{is}$ is the relevant event process, R_{is} is the associated at-risk indicator, and the w_s are weights. For these we took

$$w_s = \begin{cases} 1 & t - s \le \tau \\ \mathrm{e}^{-\rho(t-s-\tau)} & t - s > \tau \end{cases}$$

which gives equal weight to all events in the most recent τ days but discounts earlier history. After considerable experimentation we chose $\tau = 30$ and $\rho = 0.01$ for the incidence and prevalence analyses, but had no discounting for the dropout analysis.

A dynamic covariate may be on the causal pathway between a fixed covariate and the event process. The inclusion of a dynamic covariate in an analysis may therefore distort the estimation of the effects of the fixed covariates. To avoid such a distortion, at each time t we regressed each dynamic covariate on the other covariates and used the residuals from these fits as covariates when fitting the additive regression model (Fosen *et al.*, 2006). By this procedure, the estimated effects of the fixed covariates are the same in a model with dynamic covariates as in the model where only fixed covariates are included.

For the prevalence analysis we also included binary dynamic covariates which describe whether a child had diarrhoea at each of the four previous days, i.e. lags 1–4. Again we used residuals after regressing these on the fixed covariates, and in this case we also regressed each lag on the more recent values. Thus we included lag 1 in the regression model for lag 2 before defining residuals. Lags 1 and 2 were included in the model for lag 3 and so on. This helped with collinearity problems and means that the interpretation is conditional: the coefficient for lag 2 for instance measures the *extra* effect of knowing the diarrhoea status at day $t-2$ *after* allowing for known status at day $t-1$. If the diarrhoea process within an episode is Markov, there should therefore be no additional effect of knowing diarrhoea at lags >1.

6.2. Dropout analysis

For dropout analysis we fitted an additive model with all fixed and dynamic covariates included. Table 3 shows test statistics (that are standard normally distributed under the null) for assessing whether the fixed covariates are associated with dropout. It seems that older children are more likely to dropout, and perhaps people living near open sewers. People living in rain-affected accommodation and those in the lowest socio-economic category (this is defined by household income) were less likely to drop out, which is presumed to reflect willingness of the poorest people to take up a free health check. None of the dynamic covariates had any apparent effect on dropout.

Table 3. *Test statistics for covariate effects in additive regression models*

	Dynamic model		
Covariates	Dropout	Incidence	Prevalence
Male	0.12	2.78	7.50
Age			
≤ 12 months	1.48	2.80	15.35
>24 months	2.56	−13.72	−33.02
≥ 3 people/bedroom	1.16	3.71	16.11
Poor street quality	0.12		−7.36
Contaminated water storage	1.00		−4.58
Contaminated water source	−0.08	1.90	7.39
Standing water	−1.63		2.25
Open sewerage	2.05	5.58	18.72
Rain-affected accommodation	−2.05	3.70	10.13
Mother <25 yr	1.37	3.38	14.45
Low socio-economic status	−3.59		
Other children ≤ 5 yr	−1.34		

6.3. Incidence analysis

For the analysis of incidence we used backward elimination for model selection. Table 3 gives the test statistics for the selected fixed covariates. Essentially, people living in poor conditions (high density, contaminated water, open sewerage, rain-affected accommodation) have greater incidence of diarrhoea, as expected. Less experienced mothers seem to be associated with high incidence, diarrhoea episodes seem to decline with age, and there is evidence of more diarrhoea incidence amongst males than females.

Figure 4 shows a selection of cumulative regression coefficients. All covariates were included in the analysis but for space reasons we omit plots for some fixed covariates. These omitted plots are consistent with the interpretation of the test statistics in Table 3, with no evidence of time-dependent effects. The provided plot shows the cumulative baseline

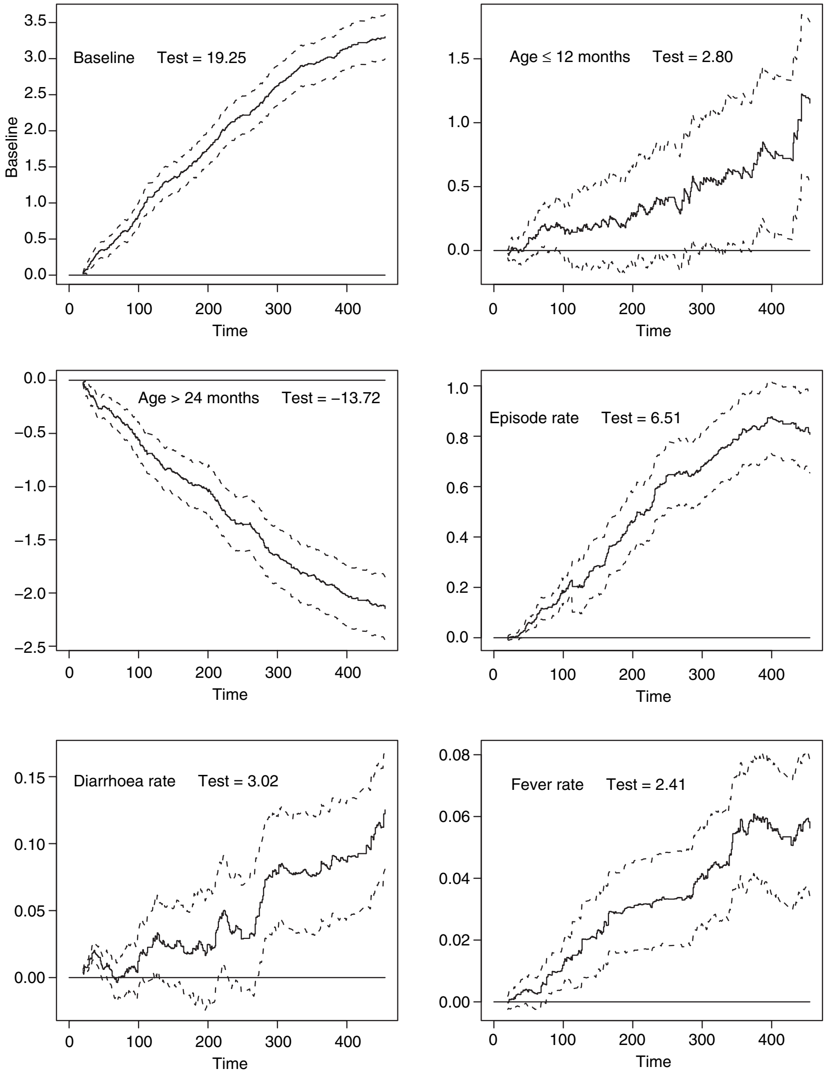

Fig. 4. Selected cumulative regression coefficients for incidence of diarrhoea, with ±2 robust standard errors.

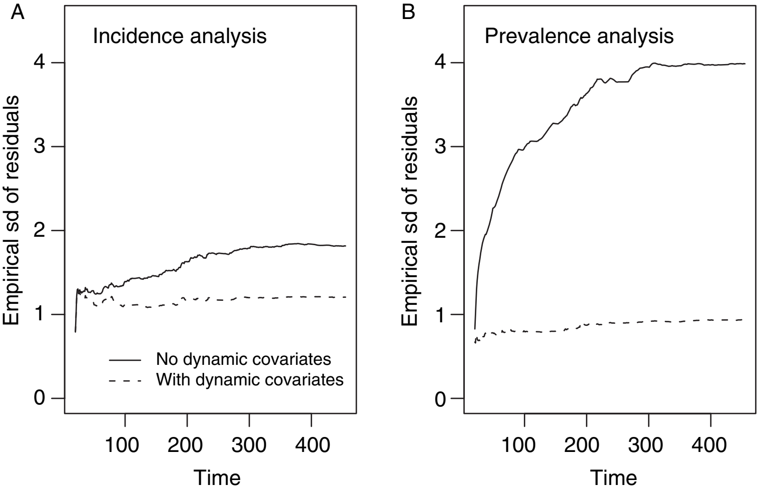

Fig. 5. Empirical standard deviations of standardized martingale residual processes.

coefficient, the two plots for categorized age, and the plots for the three dynamic covariates found to have significant effects. The plots include ±2 robust SE [cf. (11)], but not the model-based SE [cf. (9)], which were very close to the robust values. The baseline effect is fairly linear, which shows there is little evidence for the incidence rate reducing through the period of the study. The age effect is strong, with much reduced diarrhoea incidence once the child gets past about 2 years of age.

The first dynamic covariate counts the average number of previous episodes per day at risk. This is highly significant, providing evidence of a frailty effect: some children are more susceptible than others even after allowing for known risk factors. The second dynamic covariate measures the proportion of previous days on which the child had diarrhoea, and so takes into account length of episodes. Again there is a positive association, although not as strong as the episode rate. Finally, a history of fever is also predictive of future episodes. We found no evidence of interaction between dynamic covariates.

The left plot in Fig. 5 shows empirical standard deviations of the standardized martingale residual processes for incidence analyses with and without inclusion of dynamic covariates. These values should be close to 1 for a correctly specified model. Without dynamic covariates the standard deviations increase substantially as time proceeds. With dynamic covariates the pattern is stable at just over 1, suggesting the model is reasonable.

6.4. Prevalence analysis

Table 3 summarizes some of the results following our prevalence analysis, again with backward elimination for model selection. With more events and larger risk sets, the SE are smaller and more covariates are evidently statistically significant. With two exceptions, the directions of effect are positive as expected, with more diarrhoea being associated with poorer conditions. The exceptions are poor street quality and contaminated water storage, which have counter-intuitive negative association with prevalence. We suspect this is an artefact arising from near collinearity between some of the covariates.

Figure 6 gives the baseline cumulative coefficient, and those for the five dynamic covariates found to be important. The figure again has robust SE, which were once more close to the model-based ones. The dynamic covariates are the proportion of previous days with diarrhoea

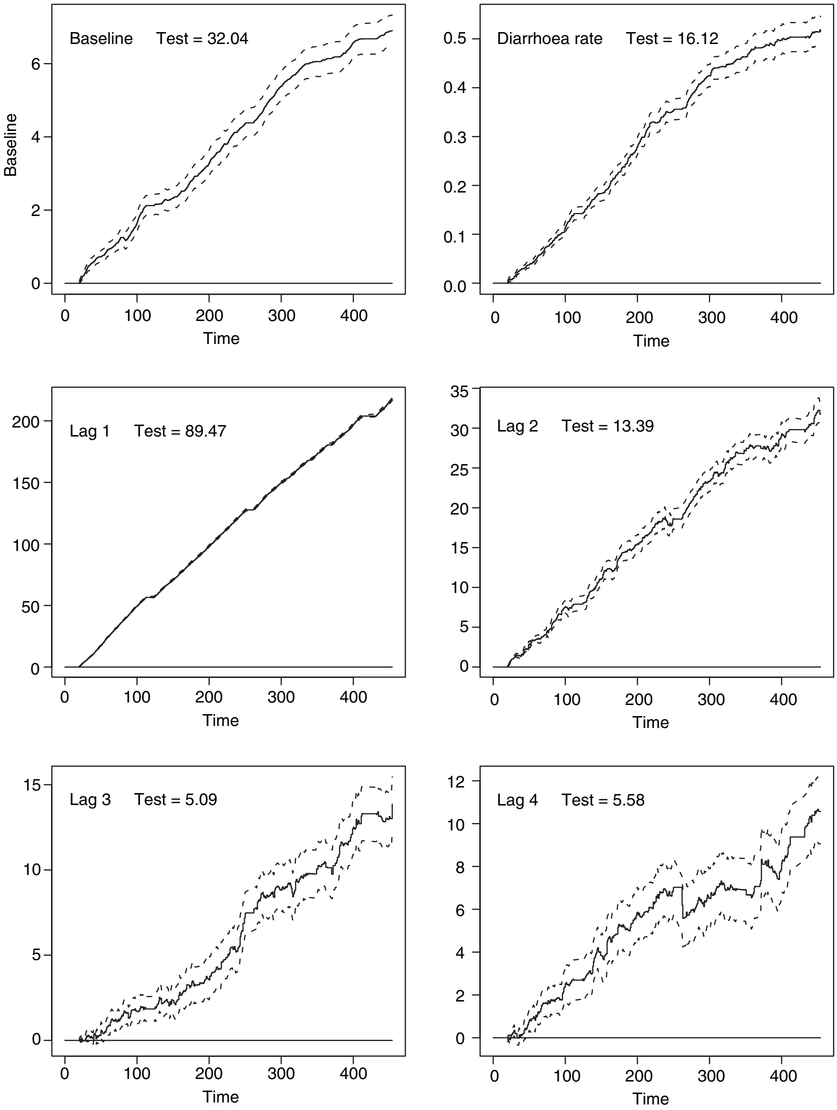

Fig. 6. Selected cumulative regression coefficients for prevalence of diarrhoea, with ±2 robust standard errors.

and the lag variables, which give the occurrence of diarrhoea d days earlier for $d = 1, 2, 3$ and 4. Note that the lag effect reduces in both magnitude and significance as d increases. Table 4 shows the estimated effects of these covariates on the probability of diarrhoea. Knowing that the child had diarrhoea the previous day increases the probability of diarrhoea by some 50%, which is close to the empirical transition probability. This is the strongest effect but note that there are still residual increases if the child was additionally known to have diarrhoea 2, 3 and 4 days earlier. The episode process is evidently not first-order Markov model.

The right-hand plot in Fig. 5 gives for the prevalence model the empirical standard deviations of the standardized martingale residual processes, with and without inclusion of dynamic covariates. The effect of including dynamic covariates is dramatic.

Table 4. *Observed and estimated probability of diarrhoea*

Diarrhoea previous days				Prevalence	
4	3	2	1	Observed (%)	Model (%)
				2	2
			✓	58	51
		✓	✓	64	60
	✓	✓	✓	66	63
✓	✓	✓	✓	72	66

The first row is unconditional, the next four assume knowledge of diarrhoea on the d immediately preceding days, for $d=1,2,3,4$.

7. Discussion

The additive modelling strategy provides a firmly based and computationally extremely efficient approach to the analysis of complex longitudinal binary data such as obtained by the Blue Bay survey. A potential disadvantage is that estimates of the conditional probabilities α_{it} are not constrained to be between 0 and 1. The possibility of negative estimates is sometimes used as an argument against using Aalen's additive model for event time data and obviously potential breaches of the upper bound of one can attract similar criticisms. For a variety of reasons we consider the advantages of the approach we have described to far outweigh these shortfalls. First, we are interested mainly in the cumulative regression functions $B_{jt}=\sum_{s=0}^{t}\beta_{jt}$, which are estimated consistently under the approach. Secondly, the powerful martingale machinery facilitated by the additive model underpins the inference, including testing and SE estimation. Thirdly, if there is interest in individual-specific prediction then it makes sense in any case to apply some local smoothing to the α_{it} to reduce noise, and this should bring estimates within the bounds.

Fourthly, and importantly, the estimation is *quick*. Each analysis of the Blue Bay data took only about 2 min, which meant that different models could be fitted and compared in real time, we could experiment with inclusion or exclusion of covariates, we could try many different weighting schemes for the dynamic covariates, and so on. Many computationally intensive methods in now standard use take hours, days or sometimes even weeks to run and genuine comparison of alternative models is not feasible. For example, we needed several days computing time to obtain the 100 bootstrap fits for the first-order Markov transition model described in section 4, using a fast programming language (Fortran). The analysis was useful, especially as it gave credence to the assumption of non-informative dropout, but nonetheless the prospect of fitting several competing models or perhaps extending beyond first-order Markov, is daunting.

There are a number of aspects to the Blue Bay data which we will consider in future work. As mentioned, 16% of children entered late. This may bring a selection effect, not so far considered in our analyses. An inverse probability weighting procedure based on entry time might be used for this investigation. Using inverse probability weighting for incidence and prevalence made rather little difference to the conclusions from the analysis, as results using unweighted least squares for estimation are similar to those using weighted least squares summarized in section 6. We suspect this may also be true for delayed entry but intend to check in further analyses. Perhaps more ambitiously, we would also like to consider the possibility of non-independence between children as it is reasonable to assume at least some of the diarrhoea to be caused by infections. The spatial locations of the children's homes are known and could be mapped, bringing the possibility of space–time modelling which we will be interested in pursuing.

Acknowledgements

We thank Agostino Strina for overseeing the methodology for data collection. The work of Ørnulf Borgan was performed partly during a sabbatical leave at the School of Mathematics and Statistics, University of Newcastle upon Tyne, UK, in the spring of 2005, and partly during a research stay at the Center for Advanced Study, Oslo, Norway, over the academic year 2005/2006. Robin Henderson's work was completed during a research stay at the same centre in spring 2006. The authors acknowledge the working facilities and inspiring environment provided there. The research of Rosemeire Fiaccone was supported in part by National Council of Technological and Scientific Development – CNPq, Brazil.

References

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Springer Lect. Notes Statist.* **2**, 1–25.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statist. Med.* **8**, 907–925.

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statist. Med.* **12**, 1569–1588.

Aalen, O. O. & Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statist. Med.* **10**, 1227–1240.

Aalen, O. O., Fosen, J., Wedon-Fekjær, H., Borgan, Ø., & Husebye, E. (2004). Dynamic analysis of multivariate failure time data. *Biometrics* **60**, 764–773.

Albert, P. S. (1991). A two-state Markov mixture model for a time series of epileptic seizure count. *Biometrics* **47**, 1371–1381.

Albert, P. S. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* **56**, 602–608.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Fosen, J., Borgan, Ø., Weedon-Fekær, H. & Aalen, O. O. (2006). Dynamic analysis of recurrent event data using the additive model. *Biom J* (in press). doi: 10.1002/bimj.200510217.

Gail, M. H., Santner, T. J. & Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **36**, 255–266.

Heagerty, P. J. & Zeger, S. L. (1998). Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *J. Amer. Statist. Assoc.* **93**, 150–162.

Hogan, J. W., Roy, J. & Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statist. Med.* **23**, 1455–1497.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The statistical analysis of failure time data,* 2nd edn. Wiley, Hoboken, NJ.

Miloslavsky, M., Keles, S. & der Laan, M. J. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *J. Roy. Statist. Soc. Ser. B* **66**, 239–257.

Oakes, D. A. (1992). Frailty models for multiple event times. In *Survival analysis: state of the art* (eds J. Klein & P. Goel). Kluwer, Dordrecht, pp. 371–379.

Scheike, T. H. (2002). The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. *Lifetime Data Anal.* **8**, 247–262.

Strina, A., Cairncross, S., Prado, M. S., Teles, C. A. & Barreto, M. L. (2005). Childhood diarrhoea symptoms, management and duration: observations from a longitudinal community study. *Trans. R. Soc. Trop. Med. Hyg.* **99**, 407–416.

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *J. Amer. Statist. Assoc.* **84**, 1065–1073.

Yue, H. & Chan, K. S. (1997). A dynamic frailty model for multivariate survival data. *Biometrics* **53**, 785–793.

Received November 2005, in final form May 2006

Ørnulf Borgan, Department of Mathematics, University of Oslo, PO Box 1053 Blindern, N-0316 Oslo, Norway.
E-mail: borgan@math.uio.no