Dept. of Math. University of Oslo
Statistical Research Report No. 9
ISSN 0806–3842 November 2005

Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil

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

Rosemeire L. Fiaccone
Department of Mathematics and Statistics
Fylde College, Lancaster University
Lancaster LA1 4YF, UK
and
Departamento de Estatistica,
Universidade Federal da Bahia, Campus Ondina,
40170-290, Salvador-BA, Brazil
e-mail: r.fiaccone@lancaster.ac.uk

Robin Henderson
School of Mathematics and Statistics
Merz Court, University of Newcastle upon Tyne
Newcastle upon Tyne NE1 7RU, UK
e-mail: Robin.Henderson@ncl.ac.uk

Mauricio L. Barreto
Instituto de Sa'ude Coletiva
Universidade Federal da Bahia, Canela
40110-170 Salvador-BA, Brazil
e-mail: mauricio@ufba.br

Abstract

This paper examines and applies methods for modelling of 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 is 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, from which daily occurrence of diarrhoea was recorded for each child being followed up. 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 and Husebye, 1991) and frailty modelling (Oakes, 1992; Yue and 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 epidemiologic study of the relationship between sanitation facilities and the occurrence of diarrhoea in children under three 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 by Albert (2000).

The data to be considered were collected during 2000-2002 in Salvador, Brazil, as part of the *Blue Bay Project*, a programme of public works and education carried out in Bahia State. Since 1997 the state government has invested over $1 billion in the project, with the aim of improving the environment and infrastructure as well as health and hygiene awareness. As part of this programme a number of surveys have been carried out, one of which forms the application to be considered in this paper.

Details of the study and data 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. An appendix provides a review of basic concepts and results on martingales and martingale transformations for discrete time longitudinal binary data.

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. The lack of environmental sanitation measures is a world-wide problem, especially in developing countries, and greatly facilitates the spread of disease. Evidence accumulated during the International Drinking Water Supply and Sanitation Decade also supports the conclusion that sanitation and water supply improvements can benefit child health (Huttly, 1994) .

Diarrhoea is the most important public health problem affected by water and sanitation and can be both waterborne and water-washed. Epidemiological evidence suggests that sanitation is at least as effective in preventing disease as improved water supply. In particular, some studies have shown the positive impact on diarrhoeal diseases after implementation of sanitation and water supply improvement programs (Victora *et al.*, 1984; Esrey and Roberts, 1991; Hoque *et al.*, 1999). In a review of more than 60 studies, Esrey *et al.* (1985) found that the median benefits of service improvements in reducing diarrhoea morbidity were 25% from improved water availability, 22% from improved excreta disposal, and 16% from water quality improvements.

In Brazil, the number of people living in accommodation connected to a sewage system increased from 35% in 1991 to 47% in 2000. The percentage of the urban population supplied with water increased from 71% in 1991 to 78% in 2000 (IBGE, 2000). However, there is a severe burden of child morbidity and mortality in some regions, in particular the Northeast (Bittencourt *et al.*, 2002; Moore *et al.*, 2000). Northeast Brazil has a population of 40 million and is the area of the country which is the poorest and has the highest mortality. In 1989, diarrhoea accounted for 25% of infant deaths. Despite a reduction in infant mortality in the last decade, diarrhoea has been the main cause of deaths in children in the Northeast region, responsible for 13.6% of deaths in children under 1 year old in 1995 (Barreto *et al.*, 1996). Moraes (1995) carried out a study in Salvador showing that children in neighbourhoods without community sanitation or with an inappropriate system had eight and three times the frequency of diarrhoea respectively as compared to those with improved sanitation.

Focusing on this topic, the Bahia state government (Brazil) and sanitation authority has implemented an extensive sanitation programme since 1997. This programme has been carried out in the metropolitan area of Salvador to improve the quality of life and the level of the health of a population estimated to be 2.3 million inhabitants. As part of this programme, the Institute of Public Health of the Federal University of Bahia has developed several studies, together called *Blue Bay*, to evaluate the impact of the resultant sanitation measures of the programme on the health of the population that inhabits the areas of operation of the sanitary sewerage system. In this paper we will focus on the morbidity of diarrhoea in children up to three 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 in details in 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 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. To illustrate further, Figure 2 shows prevalence results aggregated by month and calculated as the total days of diarrhoea divided by the total days of risk and then scaled to be days per child per year. The top plot is categorised by age of the child on entry into the study: there is a consistent decline in all groups. In the lower plot we group by actual age each day and here there is not so obvious a decline for either the youngest or oldest groups.

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 summarises these covariates. In the analysis to come all of these covariates except age are treated as binary, with the category a priori considered to bring least risk of diarrhoea coded as zero. Daily data are also available on whether or not the child had vomit or fever.

A complication for the analysis is that the children are not all observed for the full study period. Figure 3 illustrates, by showing when children were and were not observed. The figure includes only every tenth child, since resolution becomes problematic with more dense 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. Second, 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 Figure 3. 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 Figure 3. These are easily explained: the first, days 116-120 (23-27 February 2001), happened during the Salvador carnival; the second, which is slightly less obvious but is at days 237-244 (24-31 June 2001), coincided with St John's and St Paul's days, when many people took holidays;

the third, 252-258 (9-15 July 2001), happened during a strike by police; and the fourth is centred on day 421, which was Christmas Day 2001. Overall, about 20% of observations were intermittently missing.

To illustrate the data in more detail, Figure 4 shows the observation and diarrhoea pattern for ten randomly chosen children. All three types of missingness are evident in the plot. Episodes of diarrhoea tend to be relatively short, but some children are more susceptible than others. This is confirmed by the lorelogram (Heagerty and Zeger, 1998) in Figure 5, which gives the mean log odds ratios for 2×2 tables formed by presence or absence of diarrhoea on days separated by given lags. 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 zero, 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.

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 choices 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 models and parameters of interest are defined. For this situation our observations for the ith subject; $i = 1, \ldots, n$; is a binary process $\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{iT}$, where $\widetilde{Y}_{it} = 1$ if the individual experiences an event of interest at time t, $\widetilde{Y}_{it} = 0$ otherwise. For completeness we let $\widetilde{Y}_{i0} = 0$ in order to have $\widetilde{Y}_{it}$ 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 $\mathbf{x}_{it} = (x_{i1t}, \ldots, x_{ipt})^\mathsf{T}$. 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 $\widetilde{Y}_{it}$ (Kalbfleisch and 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 $\widetilde{Y}_{is}$, 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 Subsection 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 $\mathcal{H}_{it} = \mathcal{H}_{i0} \vee \sigma\{\widetilde{Y}_{i1}, \widetilde{Y}_{i2}, \ldots, \widetilde{Y}_{it}\}$. 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 $\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{iT}$ may be given by the conditional probabilities

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

A main aim for the analysis of longitudinal binary data is to study how these conditional probabilities vary over time and how they depend on covariates.

The study of the α_{it} is complicated by the 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 $\widetilde{Y}_{it}$ for subject i is observed, lost due to intermittent missingness or lost due to dropout:

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

Again, we let $Z_{i0} = 0$ in order to have $\widetilde{Y}_{it}$ defined for all $t \in \mathcal{T}$.

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

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

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 $\widetilde{Y}_{it}$ may change. It is, however, a basic assumption for our analysis that this is not the case, so that the missingness process is independent in the sense that

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

for all $t \in \mathcal{T}$, a condition similar to the independent censoring condition in event history analysis (Andersen *et al.*, 1993, Section III.2.2).

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 $\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{iT}, Z_{i1}, \ldots, Z_{iT}$ may be given by the α_{it} and the conditional missingness probabilities

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

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 $(\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{iT}, Z_{i1}, \ldots, Z_{iT})$;

$i = 1, \ldots, n$; 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

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

Note that α_{it} depends on "the past" $\mathcal{G}_{i,t-1}$ only via the covariates and $\widetilde{Y}_{i,t-1}$, 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 the value of $Z_{i,t-1}$ and adopted the multinomial logit model:

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

$l, m = 0, 1, 2, \quad y = 0, 1$. Note that the dependence between the event process $\widetilde{Y}_{it}$ and and the missingness process Z_{it} arises through the inclusion of the value of $\widetilde{Y}_{it}$ in the missingness model.

Albert proposed an EM algorithm for estimation and gave a recursive estimation procedure for calculation of the conditional probability distribution of missing $\widetilde{Y}_{it}$, 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. Since Gibbs is used, we only need to generate any missing $\widetilde{Y}_{it}$ 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 standard errors 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 characterises the dependence between the outcome $\widetilde{Y}$ 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 then turn to the additive model for the 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 additive model apply with only slight 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 also consider the filtrations $(\mathcal{F}_{it})$ 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" indicator $R_{it} = I\{Z_{it} = 0\}$ taking the value 1 if individual i is observed at time t and the value 0 otherwise, and the process $Y_{it} = R_{it}\widetilde{Y}_{it}$, registering the observed events for individual. Then

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

assuming that the all fixed and external time-varying covariates are observable.

We assume tacitly throughout that α_{it} is $(\mathcal{F}_{it})$-predictable, which implies that the dynamic time-dependent covariates in (4) are allowed to depend only on the parts of the information from $\mathcal{G}_{i,t-1}$ that are contained also in $\mathcal{F}_{i,t-1}$. Then, by (1) and (2),

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

Furthermore, unlike what was the case for the transition model, we will for the additive model assume throughout that missing is at random, so that the missingness distribution

(3) does not depend on the outcome $\widetilde{Y}_{it}$. Then

$$\lambda_{it} = \alpha_{it}\,\mathrm{E}\{P(R_{it} = 1 \,|\, \mathcal{G}_{i,t-1}) \,|\, \mathcal{F}_{i,t-1}\} = \alpha_{it}\,\pi_{it}, \tag{7}$$

where

$$\pi_{it} = P(R_{it} = 1 \,|\, \mathcal{F}_{i,t-1}) \tag{8}$$

is the conditional probability of observing $\widetilde{Y}_{it}$ given "the past" $\mathcal{F}_{i,t-1}$.

For ease of exposition we have above 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 Y_{it}. 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 use the notation $(\mathcal{F}_{it})$ also when the filtrations are enlarged.

Another comment is also in order concerning the filtrations $(\mathcal{F}_{it})$; $i = 1, \ldots, n$. 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 $\mathcal{F}_0$ generated by the $\mathcal{F}_{i0}$) we have chosen not explicitly to do so.

5.2 Inference for the additive model

In the present subsection, as well as in Subsection 5.3, we will assume that the missingness process is predictable. [From now on concepts like predictability, martingale, etc., are defined with respect to the filtrations $(\mathcal{F}_{it})$.] Then

$$\pi_{it} = P(R_{it} = 1 \,|\, \mathcal{F}_{i,t-1}) = R_{it}. \tag{9}$$

Note that the missingness process will be predictable if missingness is external to the event process or only depends on "the past". The pattern of intermittent missingness seen in Figure 3 suggest that intermittent missingness for the diarrhoea data is external, and hence predictable. It is not immediately clear that missingness due to dropouts is predictable, however, and in Subsection 5.4 we discuss this problem more closely.

By the general results for longitudinal binary data summarized in the appendix, we have the decomposition $Y_{it} = \lambda_{it} + \epsilon_{it}$ of the observation Y_{it} into a systematic part λ_{it} and a random error ϵ_{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), (7), and (9), we may write

$$Y_{it} = \beta_{0t} R_{it} + \beta_{1t}\, x_{i1t} R_{it} + \cdots + \beta_{pt}\, x_{ipt} R_{it} + \epsilon_{it}, \tag{10}$$

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 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 $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{nt})^{\mathsf{T}}$ be the vector

of observations, $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t}, \ldots, \beta_{pt})^\intercal$ the vector of regression coefficients, and $\mathbf{X}_t$ the "design matrix" with rows $\mathbf{x}_{it}^\intercal R_{it} = (1, x_{i1t}, \ldots, x_{ipt}) R_{it}$; $i = 1, \ldots, n$. Then, provided $\mathbf{X}_t$ has full rank, the least squares estimate for $\boldsymbol{\beta}_t$ becomes

$$\widehat{\boldsymbol{\beta}}_t = (\mathbf{X}_t^\intercal \mathbf{X}_t)^{-1} \mathbf{X}_t^\intercal \mathbf{Y}_t. \tag{11}$$

Let J_t be an indicator process taking the value 1 if $\mathbf{X}_t$ has full rank, and the value zero otherwise. By accumulating the least squares estimates for all times when estimation is meaningful, we obtain the estimate

$$\widehat{\mathbf{B}}_t = \sum_{s=0}^{t} J_s \, \widehat{\boldsymbol{\beta}}_s = \sum_{s=0}^{t} J_s \, (\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1} \mathbf{X}_s^\intercal \mathbf{Y}_s \tag{12}$$

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

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 $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})^\intercal$ be the vector of random errors in (10). Then $\mathbf{Y}_s = \mathbf{X}_s \, \boldsymbol{\beta}_s + \boldsymbol{\epsilon}_s$, and inserting this in (12) we obtain

$$\widehat{\mathbf{B}}_t - \mathbf{B}_t^* = \sum_{s=0}^{t} J_s \, (\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1} \mathbf{X}_s^\intercal \, \boldsymbol{\epsilon}_s.$$

Thus $\widehat{\mathbf{B}}_t - \mathbf{B}_t^*$ is a martingale transformation [cf (A.8) in the appendix], and hence a mean zero (vector-valued) martingale. In particular, $\mathrm{E}\widehat{\mathbf{B}}_t = \mathrm{E}\mathbf{B}_t^*$ for all $t \in \mathcal{T}$, so (12) is almost an unbiased estimator. By (A.10), the covariance matrix of $\widehat{\mathbf{B}}_t$ (which is approximately the same as the covariance matrix of $\widehat{\mathbf{B}}_t - \mathbf{B}_t^*$) may be estimated by

$$\widehat{\mathrm{Cov}}(\widehat{\mathbf{B}}_t) = \sum_{s=0}^{t} J_s (\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1} \mathbf{X}_s^\intercal \, \widehat{\boldsymbol{\Sigma}}_s \, \mathbf{X}_s \, (\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1}, \tag{13}$$

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

$$\widehat{\lambda}_{is} = \mathbf{x}_{it}^\intercal R_{it} \, \widehat{\boldsymbol{\beta}}_t = \{\widehat{\beta}_{0s} + \widehat{\beta}_{1s} \, x_{i1s} + \cdots + \widehat{\beta}_{ps} \, x_{ips}\} \, R_{is} \tag{14}$$

a model based estimate of λ_{it}; cf. (4), (7), and (9). Moreover, by the martingale central limit theorem, (12) is approximately multivariate normally distributed in large samples.

The above derivations are similar to those for Aalen's additive model for continuous time event history data, see Section VII.4 in Andersen *et al.* (1993) for a review. Also a test for the hypothesis

$$\mathrm{H}_{0j} \, : \, \beta_{jt} = 0 \quad \text{for all } t \in \mathcal{T},$$

i.e. that the jth covariate has no effect, can be derived in a similar manner as for Aalen's additive model. A test for H_{0j} may be based on a test statistic of the form

$$U_j = \sum_{s \in \mathcal{T}} L_{js} \widehat{\beta}_{js}, \tag{15}$$

where L_{js} is a predictable weight process. Following Aalen (1989) we will let L_{js} be the reciprocal of the $(j+1)$st diagonal element of the matrix $(\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1}$. Under H_{0j} the test

statistic (15) is a martingale transformation of the form (A.8) with $\mathbf{K}_s = J_s \mathbf{L}_s^{(j)} (\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1} \mathbf{X}_s^\intercal$, where $\mathbf{L}_s^{(j)}$ is the $(p+1)\times(p+1)$ matrix with all entries equal to zero except the $(j+1)$st diagonal element which equals L_{js}. A variance estimator, $\widehat{\mathrm{var}}(U_j)$, of (15) is given by (A.10) in the appendix, evaluated at $t = T$, with $\widehat{\lambda}_{is}$ given by (14). By the martingale central limit theorem we may then conclude that the standardised test statistic $U_j / \{\widehat{\mathrm{var}}(U_j)\}^{1/2}$ is approximately standard normal when H_{0j} holds true and sample size is reasonable.

The estimator (13) of the covariance matrix of $\widehat{\mathbf{B}}_t$ is valid when our model for $\lambda_{it} = P(Y_{it} = 1 \,|\, \mathcal{F}_{i,t-1})$ 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

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

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

$$\widetilde{\mathrm{Cov}}(\widehat{\mathbf{B}}_t) = \sum_{i=1}^{n} \mathbf{Q}_{it}^{\otimes 2} \tag{16}$$

for the covariance matrix of (12). Here

$$\mathbf{Q}_{it} = \sum_{s=0}^{t} J_s (\mathbf{X}_s^\intercal \mathbf{X}_s)^{-1} \mathbf{x}_{is} (Y_{is} - \widehat{\lambda}_{is}),$$

where for a vector $\mathbf{a}$, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\intercal$.

5.3 Martingale residual processes

One important tool to assess the fit of an additive model, is the martingale residual processes. These were introduced by Aalen (1993) in the context of his additive model for survival and event history data (Aalen, 1980, 1989), and their use for continuous time recurrent event data was illustrated by Aalen *et al.* (2004). We will here 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

$$M_{it} = N_{it} - \Lambda_{it} \tag{17}$$

is a martingale; cf. the appendix. The idea is now to replace the Λ_{it} in (17) by its estimate $\widehat{\Lambda}_{it} = \sum_{s=0}^{t} \widehat{\lambda}_{is}$ under the model [cf. (14)] to obtain the martingale residual process $\widehat{M}_{it}$. If the model is correctly specified, each of the n individual residual processes should behave as a martingale.

More specifically, we introduce the vector $\widehat{\boldsymbol{\Lambda}}_t = (\widehat{\Lambda}_{1t}, \ldots, \widehat{\Lambda}_{nt})^\intercal$, and note that by (11) and (14), this may be given as

$$\widehat{\boldsymbol{\Lambda}}_t = \sum_{s=0}^{t} J_s \, \mathbf{X}_s \, \widehat{\boldsymbol{\beta}}_s = \sum_{s=0}^{t} J_s \, \mathbf{H}_s \, \mathbf{Y}_s,$$

where

$$\mathbf{H}_s = \mathbf{X}_s \left(\mathbf{X}_s^\intercal \mathbf{X}_s\right)^{-1} \mathbf{X}_s^\intercal$$

is the "hat matrix". Then if we introduce the vector $\mathbf{N}_t = \sum_{s=0}^{t} J_s \mathbf{Y}_s$ of counting processes, restricted to time points where estimation is possible, the vector $\widehat{\mathbf{M}}_t = (\widehat{M}_{1t}, \ldots, \widehat{M}_{nt})^\intercal$ of martingale residual processes may be written

$$\widehat{\mathbf{M}}_t = \mathbf{N}_t - \widehat{\mathbf{\Lambda}}_t = \sum_{s=1}^{t} J_s(\mathbf{I} - \mathbf{H}_s)\mathbf{Y}_s. \tag{18}$$

When the additive model is correctly specified, $\mathbf{Y}_s = \mathbf{X}_s \boldsymbol{\beta}_s + \boldsymbol{\epsilon}_s$, and the vector of martingale residual processes becomes

$$\widehat{\mathbf{M}}_t = \sum_{s=1}^{t} J_s(\mathbf{I} - \mathbf{H}_s)\boldsymbol{\epsilon}_s,$$

i.e., it is a martingale transformation, and hence a mean zero martingale; cf. (A.8). By (A.10) the covariance matrix of the vector of martingale residual processes may be estimated by

$$\widehat{\mathrm{Cov}}(\widehat{\mathbf{M}}_t) = \sum_{s=0}^{t} J_s(\mathbf{I} - \mathbf{H}_s)\, \widehat{\mathbf{\Sigma}}_s\, (\mathbf{I} - \mathbf{H}_s)^\intercal, \tag{19}$$

where $\widehat{\mathbf{\Sigma}}_s$ is given just above (14).

From the vector of martingale residual processes and its estimated covariance matrix, we may derive standardised martingale residual processes by dividing each entry of (18) by the square root of the corresponding diagonal element of (19). If the model is correctly specified, these should have mean zero and variance one. Following Fosen *et al.* (2004) we will in Subsections 6.3 and 6.4 check the fit of a model by plotting the empirical standard deviation of the standardised residual processes as a function of time. If a model fits reasonably well, the empirical standard deviation should be about one, while larger values indicate a poor fitting model.

5.4 Dealing with dropouts

Although we are convinced that intermittent missingness in the Blue Bay data is essentially administrative, and hence predictable [cf. (9)], it is not clear that the same can be said for dropout. If dropout is associated with response history then a selection effect may lead to biased results. However, this can be adjusted for by the use of inverse probability weighting, as described by, for instance, Robins *et al.* (1995).

To describe in more detail how this can be done, we first consider the situation where the probabilities π_{it} [cf. (8)] are assumed known. Then we may for each $t \in \mathcal{T}$ estimate the regression coefficients β_{jt} of the additive model (4) by regressing the $\widetilde{Y}_{it}$ on the covariates x_{ijt} using *weighted* least squares with weights R_{it}/π_{it}. The $\widetilde{Y}_{it}$ are not all observed, but by interpreting $0/0$ as 0, the use of the weights R_{it}/π_{it} ensures that only the $\widetilde{Y}_{it}$ that actually are observed enter into the estimation procedure. The weighted least squares gives rise to the estimating equations (with $x_{i0t} = 1$)

$$\sum_{i=1}^{n} \frac{R_{it}}{\pi_{it}} \left(\widetilde{Y}_{it} - \alpha_{it}\right) x_{ijt} = 0 \qquad\qquad j = 0, 1, \ldots, p \tag{20}$$

where the regression coefficients enter into the estimating equations via the additive model for α_{it} [cf. (4)]. By the relation $Y_{it} = R_{it}\widetilde{Y}_{it}$ and the predictability of π_{it} and the covariates, we get using (6), (7), and (8)

$$\mathrm{E}\left\{\frac{R_{it}}{\pi_{it}}(\widetilde{Y}_{it} - \alpha_{it})x_{ijt} \,|\, \mathcal{F}_{i,t-1}\right\} = x_{ijt}\left\{\frac{\mathrm{E}(Y_{it} \,|\, \mathcal{F}_{i,t-1})}{\pi_{it}} - \alpha_{it}\frac{\mathrm{E}(R_{it} \,|\, \mathcal{F}_{i,t-1})}{\pi_{it}}\right\} = 0.$$

It thus follows that the estimating equations (20) are (conditionally) unbiased, and this justifies the use of the weighted least squares procedure with weights R_{it}/π_{it} when the π_{it} are assumed known.

In practice, the π_{it} will have to be estimated, and this leads to a two-stage procedure. At the first stage dropout is the (single) event of interest and we fit an additive model for the dropout probabilities $1 - \pi_{it}$ including all fixed and dynamic covariates. This leads, at each time point t, to an estimated (conditional) probability $\widehat{\pi}_{it}$ for subject i still being in the study. Then, at the second stage, the $R_{it}/\widehat{\pi}_{it}$ are built in as weights in the incidence or prevalence analysis. Inclusion of the weights $R_{it}/\widehat{\pi}_{it}$ requires some modifications to the methods given Subsections 5.2 and 5.3, essentially we need to scale both the vector of responses $\mathbf{Y}_t$ and the "design matrix" $\mathbf{X}_t$ by the square roots of the weights.

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.

6.1 Fixed and dynamic covariates

Table 1 summarises 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 categorisation 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 categorised 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

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

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

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

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. In order to avoid such a distortion we will at each time t regress each dynamic covariate on the other covariates and use the residuals from these fits as covariates when fitting the additive regression model (Fosen *et al.*, 2004). 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 greater than one.

6.2 Dropout analysis

Since the primary purpose of the dropout analysis is to provide adjustment weights for the incidence and prevalence analyses, all fixed and dynamic covariates were included in the additive regression model for dropout. Table 3 shows standardised test statistics, derived as described below (15), 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.

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 the poorest conditions have greater incidence of diarrhoea, as expected. More experienced mothers seem to be associated with lower incidence, and diarrhoea episodes are more common in very young infants.

Figure 6 shows the cumulative baseline coefficient, the two plots for categorised age, and the plots for the three dynamic covariates found to have significant effects. The plots include ± 2 robust standard errors [cf. (16)], but not the model-based standard errors [cf. (13)], 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 two 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, though 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.

Figure 7 shows empirical standard deviations of the standardised martingale residual processes for incidence analyses with and without inclusion of dynamic covariates. These values should be close to one 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 one, suggesting the model is reasonable.

6.4 Prevalence analysis

Table 3 summarises some of the results following our prevalence analysis, again with backward elimination for model selection. With more events and larger risk sets, the standard errors are smaller and more covariates are evidently statistically significant. With two exceptions, the directions of effect are as expected, the exceptions being 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 8 gives the baseline cumulative coefficient, and those for the five dynamic covariates found to be important, again with robust standard errors, which were once more close to model-based ones. The dynamic covariates are the proportion of previous days with diarrhoea 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 that 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. It is worth mentioning here another advantage of the additive modelling approach: we can investigate the effect of one group of covariates (e.g. the lags) without specifying the values of the others. Table 4 assumes reference (zero) values but the estimated lag effect is the same for all combinations.

Figure 9 gives for the prevalence model the empirical standard deviations of the standardised martingale residual processes, with and without inclusion of dynamic covariates. The effect of including dynamic covariates is dramatic.

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 zero and one. 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. Second, the powerful martingale machinery facilitated by the additive model underpins the inference, including testing and standard error estimation. Third, 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.

Fourth, and importantly, the estimation is *quick*. Each analysis of the Blue Bay data took only about two minutes, 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 intermittent missingness, 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 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 mimicking that used for dropout, but based on entry time, might be used for this investigation. Using the inverse probability method for dropout actually 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 summarised 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 in subsequent work.

Appendix: Basic results for longitudinal binary data

In this appendix we summarise some results for longitudinal binary data in discrete time that are needed in Section 5. The results follow by standard results for time-discrete martingales as given, e.g., in Williams (1991).

We start out by considering longitudinal binary data for a generic individual. Thus assume that the binary stochastic process $Y = (Y_t)$; $t \in \mathcal{T}$; with $Y_0 = 0$, is adapted to a filtration $(\mathcal{F}_t)$, and let N be the process $N_t = \sum_{s=0}^{t} Y_s$ counting the number of 1s in Y up to and including time t. We introduce the process

$$\lambda_t = P(Y_t = 1 \,|\, \mathcal{F}_{t-1}) = \mathrm{E}(Y_t \,|\, \mathcal{F}_{t-1}), \tag{A.1}$$

as well as its cumulative counterpart $\Lambda_t = \sum_{s=0}^{t} \lambda_s$, and note that the processes λ and Λ are predictable (i.e. λ_t and Λ_t are $\mathcal{F}_{t-1}$-measurable for each $t \in \mathcal{T}$). We then consider the

process $M = N - \Lambda$. Note that $M_0 = 0$, and that

$$M_t = N_t - \Lambda_t = \sum_{s=0}^{t} \epsilon_s,$$

where $\epsilon_t = Y_t - \lambda_t$. Using (A.1), we see that $\mathrm{E}(\epsilon_t \,|\, \mathcal{F}_{t-1}) = 0$. Thus

$$\mathrm{E}(M_t \,|\, \mathcal{F}_{t-1}) = \mathrm{E}(M_{t-1} + \epsilon_t \,|\, \mathcal{F}_{t-1}) = M_{t-1} + \mathrm{E}(\epsilon_t \,|\, \mathcal{F}_{t-1}) = M_{t-1},$$

which shows that M is a martingale. In particular $\mathrm{E}M_t = \mathrm{E}\mathrm{E}(M_t \,|\, \mathcal{F}_0) = M_0 = 0$, for all $t \in \mathcal{T}$, so M has mean zero. The predictable variation process $\langle M \rangle$ of M is given by

$$\langle M \rangle_t = \sum_{s=0}^{t} \mathrm{Var}(\epsilon_s \,|\, \mathcal{F}_{s-1}) = \sum_{s=0}^{t} \lambda_s(1 - \lambda_s). \tag{A.2}$$

The process $M^2 - \langle M \rangle$ is a mean zero martingale. In particular, $\mathrm{Var}M_t = \mathrm{E}(M_t^2) = \mathrm{E}\langle M \rangle_t$.

If $K = \{K_t\}$ is a predictable process, we may define a new process $K \circ M$ by

$$(K \circ M)_t = \sum_{s=0}^{t} K_s \, (M_s - M_{s-1}) = \sum_{s=0}^{t} K_s \, \epsilon_s. \tag{A.3}$$

This process, which is a discrete version of a stochastic integral, is denoted the *transformation* of M by K. It is easy to check, using the predictability of K, that $K \circ M$ is a martingale. Further its predictable variation process is given by

$$\langle K \circ M \rangle_t = \sum_{s=0}^{t} \mathrm{Var}(K_s \, \epsilon_s \,|\, \mathcal{F}_{s-1}) = \sum_{s=0}^{t} K_s^2 \, \lambda_s(1 - \lambda_s). \tag{A.4}$$

Assume then that we have n binary time series Y_i; $i = 1, \ldots, n$; and write N_i, λ_i, M_i, and ϵ_i for the processes derived from these. We assume that the processes are adapted or predictable as described above with respect to a common filtration $(\mathcal{F}_t)$. For the discrete time situation considered here, we may have $Y_{it} = 1$ for two or more indices i with positive probability. Thus the counting processes N_i may have common jumps, and therefore the n-variate process $(N_1, \ldots, N_n)$ is *not* a multivariate counting process. However, many of the results from the theory of continuous time counting processes carry over to the present situation if we assume that the individual errors ϵ_{it} are conditionally uncorrelated, specifically that for all t and all $i \neq j$ we have

$$\mathrm{Cov}(\epsilon_{it}, \epsilon_{jt} \,|\, \mathcal{F}_{t-1}) = 0. \tag{A.5}$$

For then the predictable covariance processes $\langle M_i, M_j \rangle$ become

$$\langle M_i, M_j \rangle_t = \sum_{s=0}^{t} \mathrm{Cov}(\epsilon_{is}, \epsilon_{js} \,|\, \mathcal{F}_{s-1}) = \delta_{ij} \, \langle M_i \rangle_t, \tag{A.6}$$

with δ_{ij} a Kronecker delta. Thus the martingales M_i are orthogonal; a key property for the "classical" continous time counting process theory. We note that the assumption (A.5) is fulfilled when the binary stochastic processes Y_i; $i = 1, \ldots, n$; are conditionally independent given $\mathcal{F}_0$, as is the case in the main body of the paper.

If $K^{(1)} = \{K_t^{(1)}\}$ and If $K^{(2)} = \{K_t^{(2)}\}$ are predictable processes, it follows by (A.2) and (A.5) that

$$\begin{aligned} \langle K^{(1)} \circ M_i, K^{(2)} \circ M_j \rangle_t &= \sum_{s=0}^{t} \mathrm{Cov}(K_s^{(1)}\, \epsilon_{is}, K_s^{(2)}\, \epsilon_{js} \,|\, \mathcal{F}_{s-1}) \\ &= \delta_{ij} K_s^{(1)}\, K_s^{(2)}\, \lambda_{is}(1-\lambda_{is}). \end{aligned} \tag{A.7}$$

We now introduce $\mathbf{M}_t = (M_{1t}, \ldots, M_{nt})^\intercal$ and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})^\intercal$; the vectors of the martingales and their increments, and note that the predictable variation process $\langle \mathbf{M} \rangle$ of $\mathbf{M}$ is the matrix valued processes with (i,j)th entry equal to $\langle M_i, M_j \rangle$. Further we let $\mathbf{K} = \{\mathbf{K}_t\} = \{K_{jlt}\}$ be a $p \times n$ matrix of predictable processes. Then the transformation of $\mathbf{M}$ by $\mathbf{K}$ is the p-dimensional mean zero vector-valued martingale $\mathbf{K} \circ \mathbf{M}$ given by

$$(\mathbf{K} \circ \mathbf{M})_t = \sum_{s=0}^{t} \mathbf{K}_s\, \boldsymbol{\epsilon}_s. \tag{A.8}$$

Using (A.4), (A.6), and (A.7), we find that the $p \times p$ matrix-valued predictable variation process $\langle \mathbf{K} \circ \mathbf{M} \rangle$ of $\mathbf{K} \circ \mathbf{M}$ is given by

$$\langle \mathbf{K} \circ \mathbf{M} \rangle_t = \sum_{s=0}^{t} \mathbf{K}_s\, \boldsymbol{\Sigma}_s\, \mathbf{K}_s^\intercal \tag{A.9}$$

where $\boldsymbol{\Sigma}_s = \mathrm{diag}\{\lambda_{is}(1-\lambda_{is})\}$ is the $n \times n$ diagonal matrix with ith diagonal element equal to $\lambda_{is}(1-\lambda_{is})$. In particular the covariance matrix of $(\mathbf{H} \circ \mathbf{M})_t$ takes the form $\mathrm{Cov}(\mathbf{K} \circ \mathbf{M})_t = \sum_{s=0}^{t} \mathrm{E}\,(\mathbf{K}_s\, \boldsymbol{\Sigma}_s\, \mathbf{K}_s^\intercal)$, and it may be estimated by

$$\widehat{\mathrm{Cov}}(\mathbf{K} \circ \mathbf{M})_t = \sum_{s=0}^{t} \mathbf{K}_s\, \widehat{\boldsymbol{\Sigma}}_s\, \mathbf{K}_s^\intercal \tag{A.10}$$

where $\widehat{\boldsymbol{\Sigma}}_s = \mathrm{diag}\{\widehat{\lambda}_{is}(1-\widehat{\lambda}_{is})\}$ with $\widehat{\lambda}_{is}$ an estimator of (A.1).

Acknowledgements

We thank Agostino Strina for overseeing the methodology for data collection. The work of Ørnulf Borgan was done 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, the academic year 2005/2006. Both institutions are acknowledged for providing the best working facilities and an inspiring environment for research. The research of R. Fiaccone was supported in part by National Counsel of Technological and Scientific Development - CNPq, Brazil.

References

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

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907-925.

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

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

Aalen, O. O. and Husebye, E. (1991). Statistical analysis of repeated events forming reneval processes. *Statistics in Medicine* **10**, 1227–1240.

Albert, P. S. (1991). A two-stage Markov mixture model for a time series of epileptic seizure count. *Biometrics* **47**, 1371–1240.

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. and Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Barreto, M. L., Carmo, E., Santos, C. and Ferreira, L. (1996). "Emergentes", "re-emergentes" e "permanecentes": Tendancias recentes das doenas infecciosas e parasitarias no Brasil (in Portuguese). *Informe Epidemiolgico SUS* **3**, 7-18.

Bittencourt, S. A., Leal, M. C. and Santos, M. O. (2002). Hospitalizaes por diarria infecciosa no estado do Rio de Janeiro (in Portuguese). *Caderno de Saude Publica* **18**, 747-754.

Esrey, S. A., Feachem, R. and Hughes, J. M. (1985). Interventions for the control diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. *Bulletin of the World Health Organisation* **63**, 757–772.

Esrey, S. A.and Potash, J. B. and Roberts, L. (1991). Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, shistosomiasis, ad trachoma. *Bulletin of the World Health Organisation* **69**, 609–621.

Fosen, J., Borgan, Ø., Weedon-Fekær, H. and Aalen, O. O. (2004). Path analysis for survival data with recurrent events. Statistical Research Report no 9/2004 (`http://www.math.uio.no/eprint/stat_report/2004/09-04.html`). Department of Mathematics, University of Oslo.

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

Heagerty, P. J. and Zeger, S. L. (1998). Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *Journal American Statistical Association* **93**, 150–162.

Hoque, B., Chakraborty, J. and Chowdhury, J. (1999). Effects of environmental factors on child survival in Bangladesh: a case control study. *Public Health* **113**, 57–64.

Huttly, S. R. (1994). Water, sanitation and health in developing countries. Pp. 251–258 in *Water and public health* (eds. ds: Golding, Noah, and Stanwell-Smith). Smith Gordon, London.

IBGE - Brazilian Instutute of Geography and Statistics (2000). PNAD: Pesquisa nacional por amostra de domicilios (in Portuguese) (`www.ibge.org.br/home/estatisticas/pesquisas`).

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data. 2nd ed.*. Wiley, Hoboken, New Jersey.

Miloslavsky, M., Keles, S. and van der Laan, M. J. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal Royal Statistical Society* **B 66**, 239–257.

Moore, S. R., Lima, A. A. and Schorling, J. B. (2000). Changes over time in the epidemiology of diarrhea and malnutrition among children in an urban Brazilian shantytown, 1989 to 1996. *International Journal Infectious Disease* **4**, 179–186.

Moraes, L. R. S. (1995). Health impact of drainage and sewerage on poor urban areas in Salvador, Brazil. PhD thesis, London School of Hygiene and Tropical Medicine.

Oakes, D A. (1992). Frailty models for multiple event times. Pp.371–379 in *Survival Analysis: State of the Art* (eds. Klein, J. and Goel, P.). Kluwer, Dordreccht.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis fo semiparametric regression models for repeated outcomes in the presence if missing data. *Journal of the American Statistical Association* **90**, 106–121.

Scheike, T. H. (2002). The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. *Lifetime Data Analysis* **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. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **99**, 407–416.

Victora, C. G., Smith, P. G. and Vaughan, J. P. (1984). Water supply, sanitation and housing in relation to the risk of infant mortality from diarrhoea. *International Journal of Epidemiology* **17**, 651–654.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal American Statistical Association* **84**, 1065–1073.

Williams, D. (1991). *Probability with martingales*. Cambridge University Press, Cambridge (UK).

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

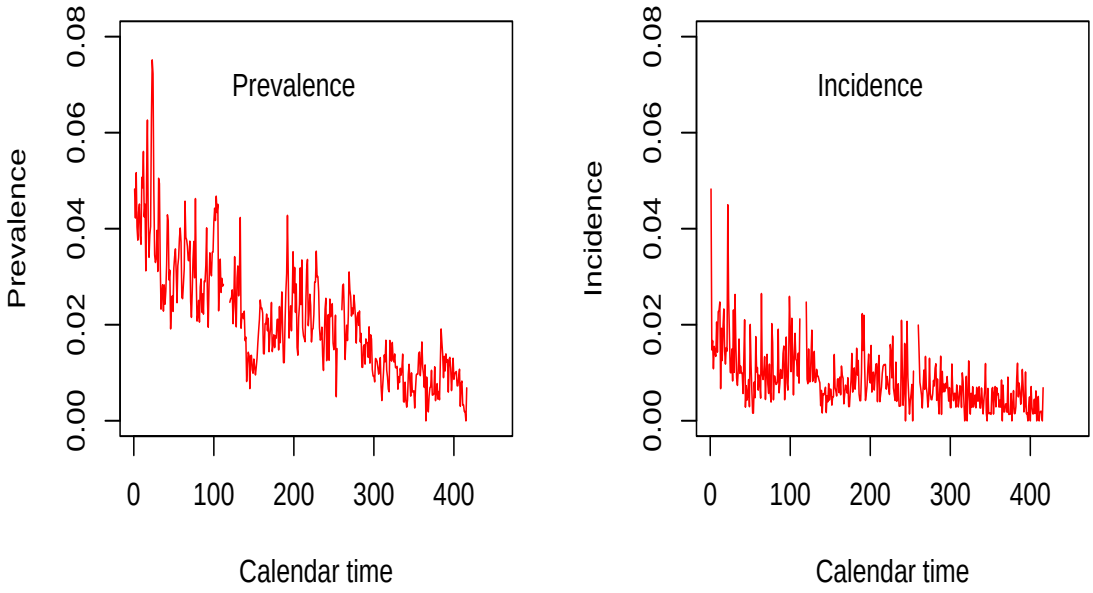
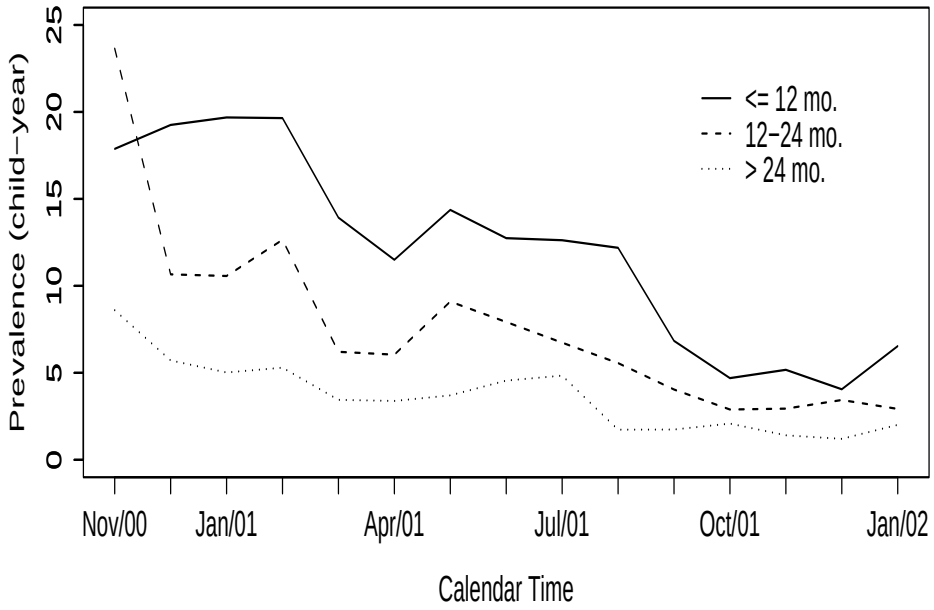


Figure 1: Prevalence and incidence of diarrhoea after start of study.



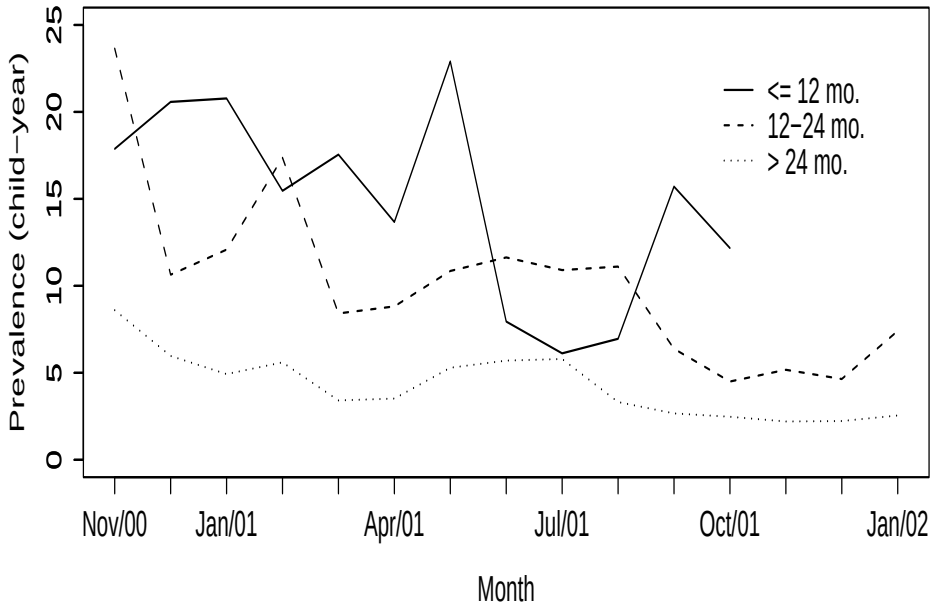


Figure 2: Prevalence (cases per child-year) by age of child on entry (top plot) and current age (bottom plot), with monthly aggregation.

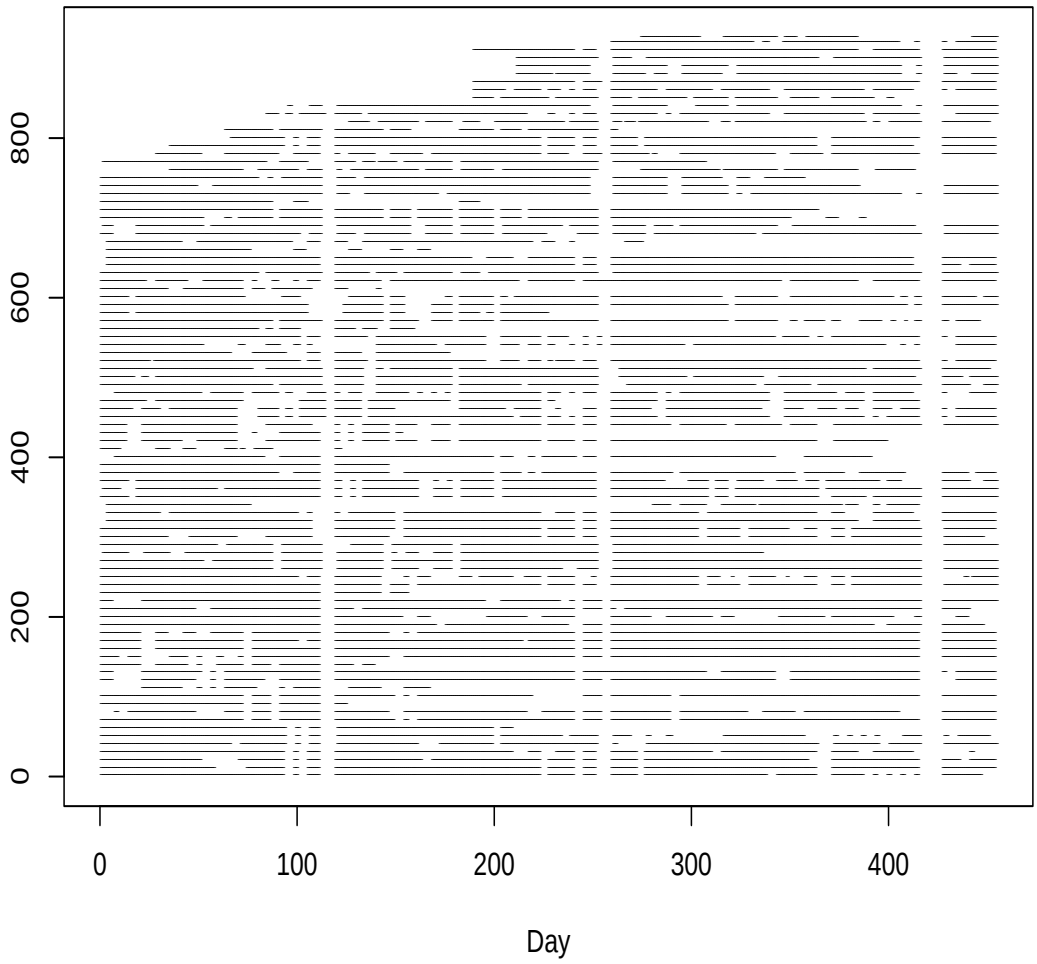


Figure 3: Observation pattern. Horizontal lines indicate where data for each child is available, for every tenth child only.

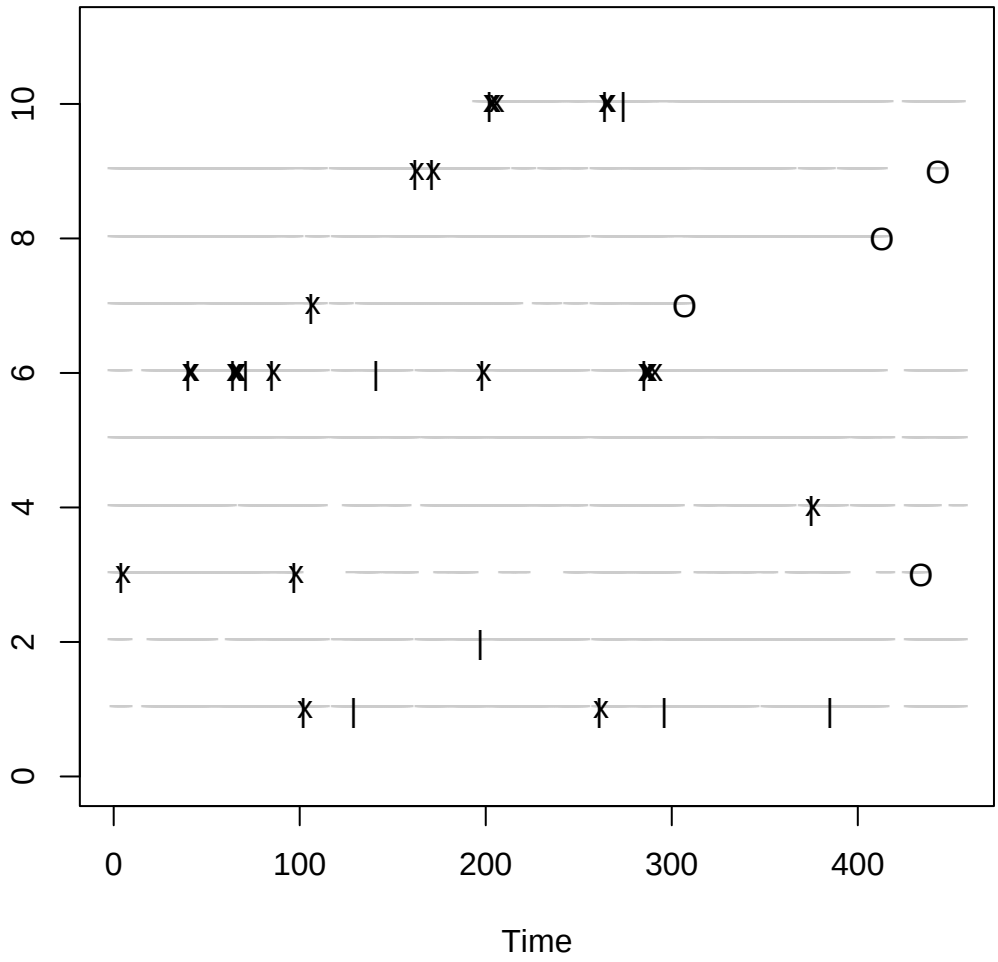


Figure 4: Specimen data. One line for each of 10 randomly chosen children. Grey lines indicate where data is available; vertical bars mark the starts of diarrhoea episodes; crosses mark subsequent days with diarrhoea; circles indicate dropout.

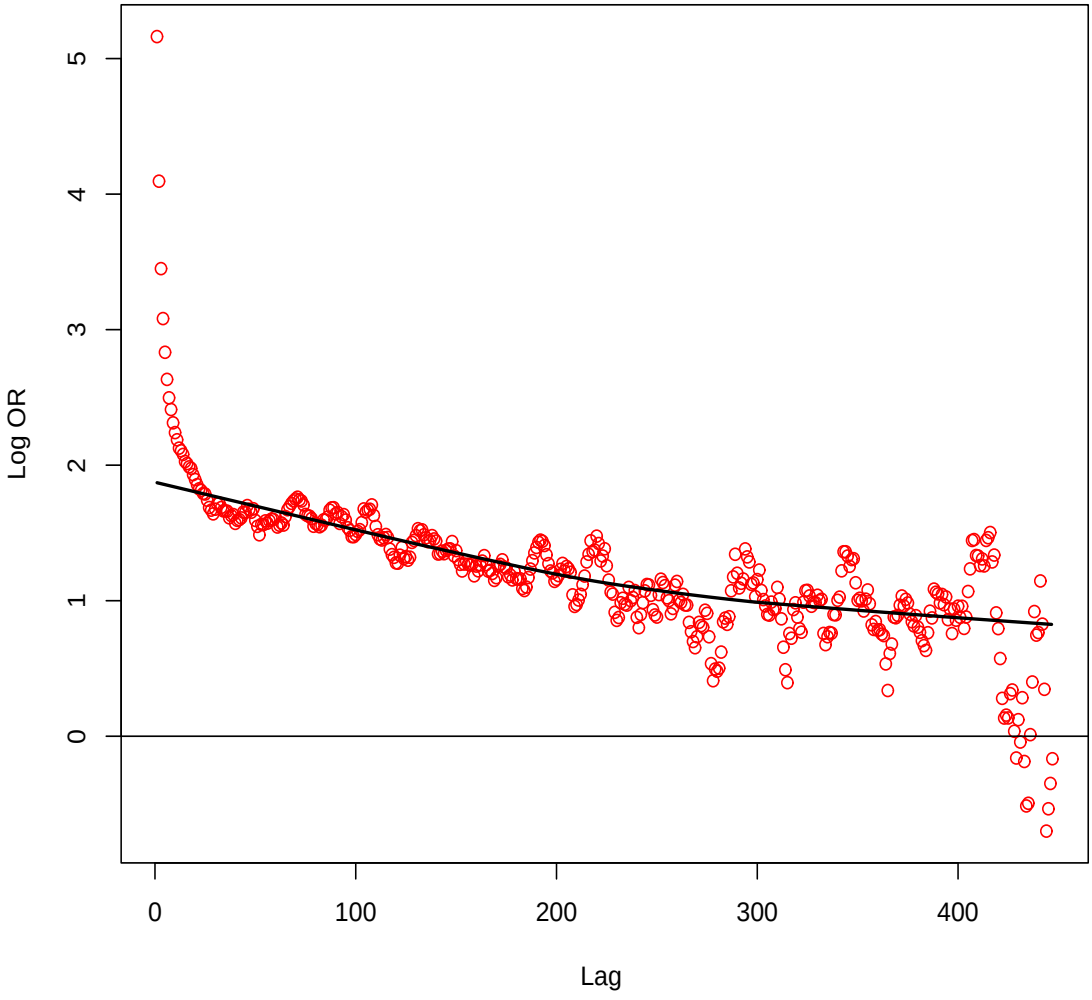


Figure 5: Lorelogram for diarrhoea data: see text for explanation.

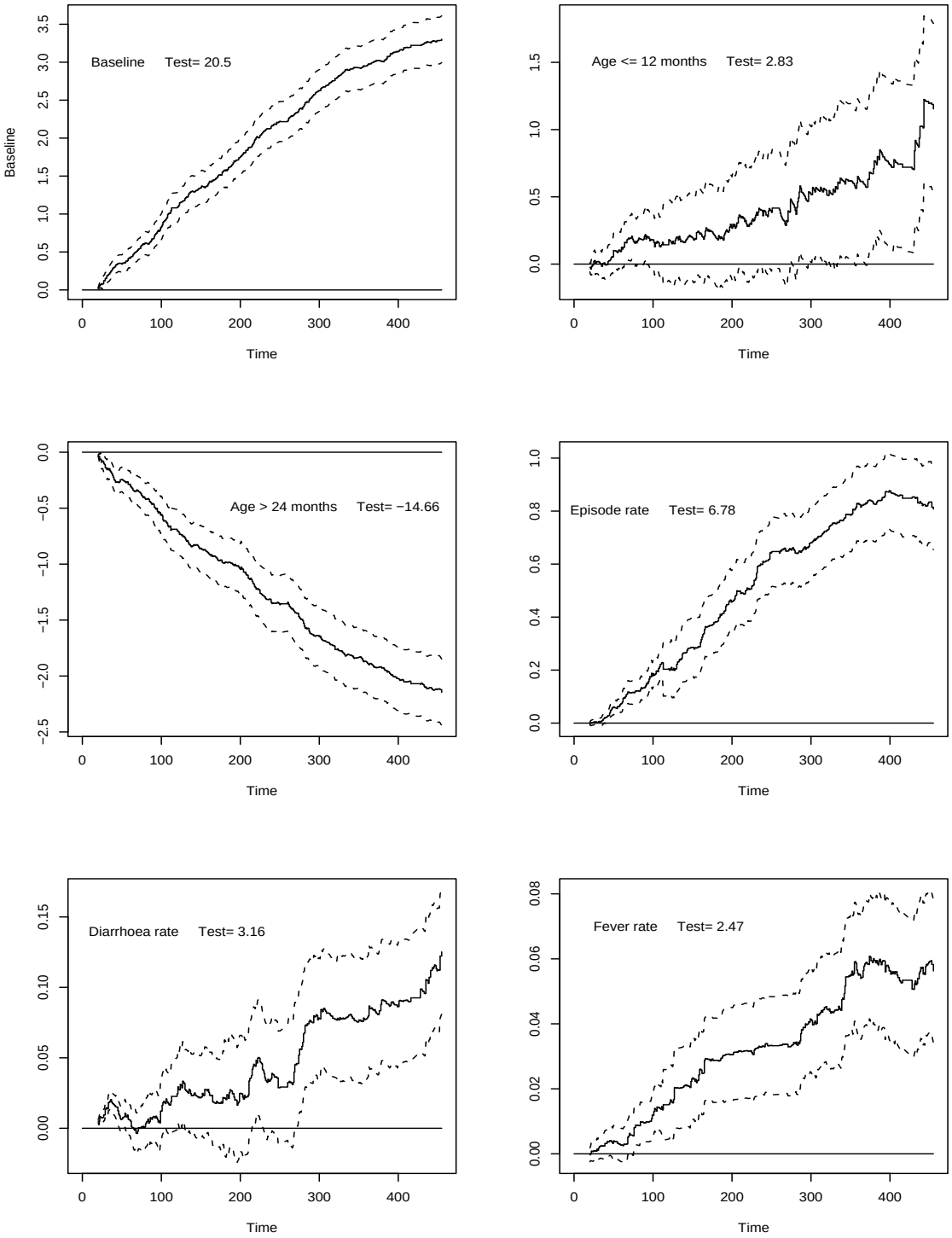


Figure 6: Selected cumulative regression coefficients for incidence model, with ±2 robust standard errors.

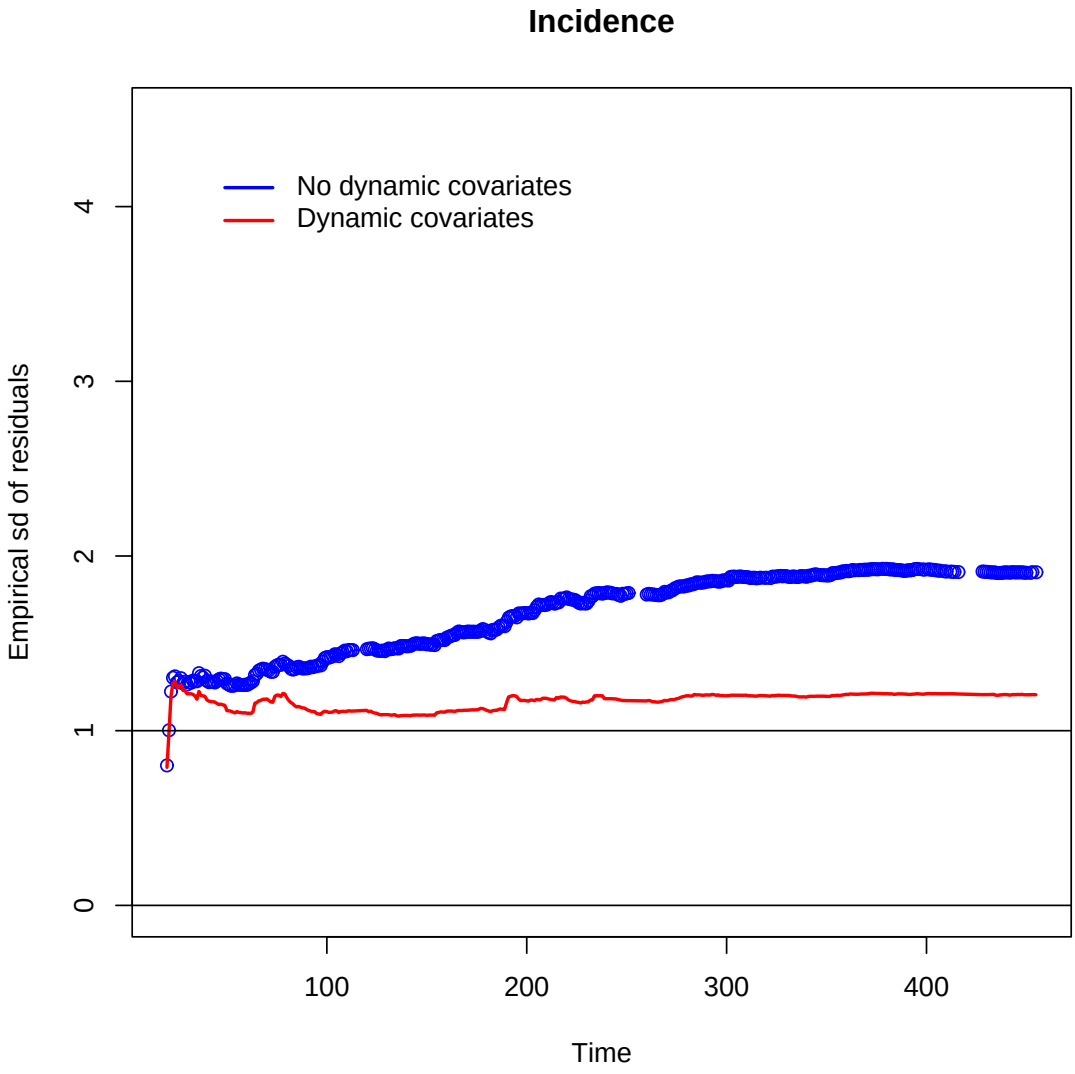


Figure 7: Incidence model: empirical standard deviations of standardised martingale residual processes.

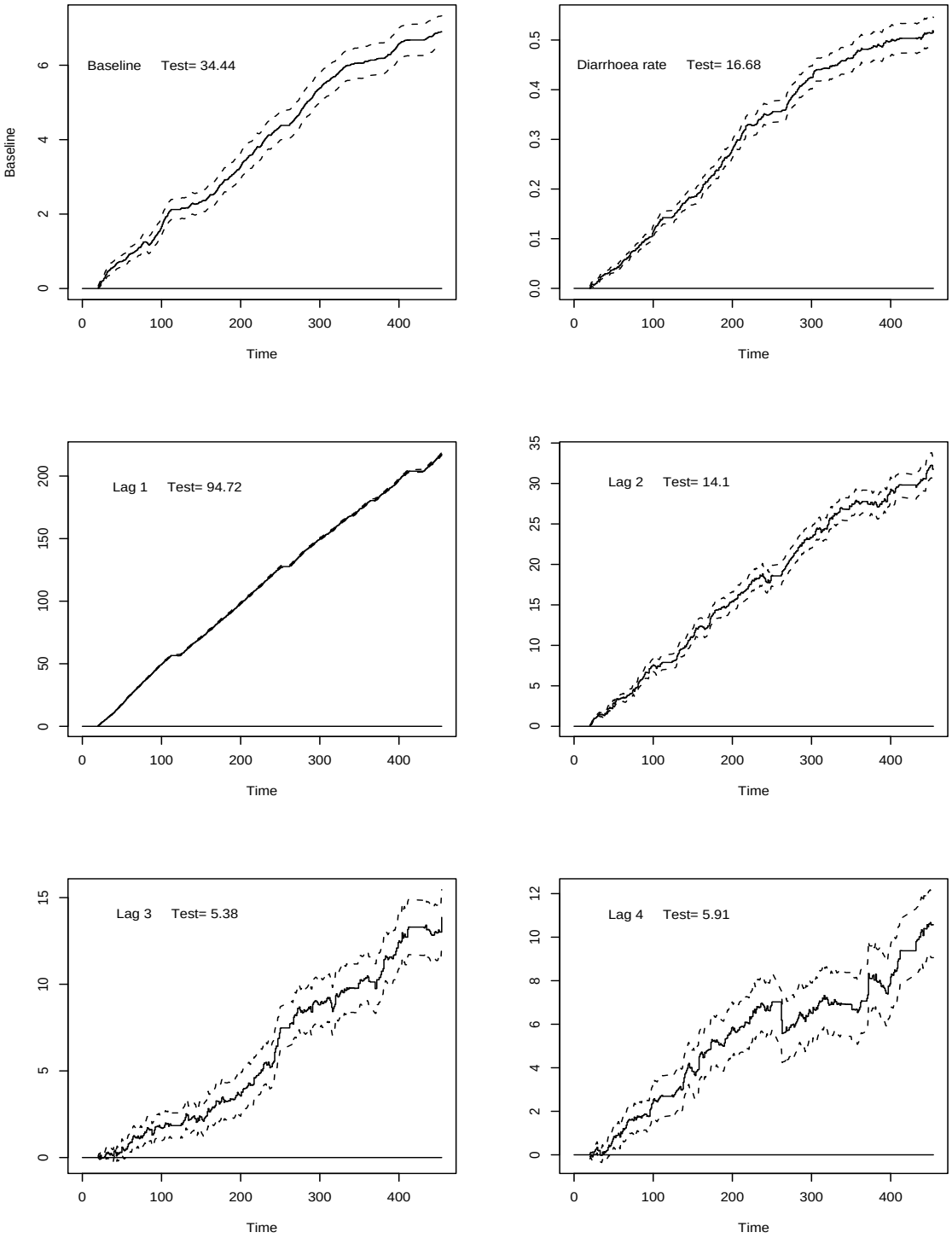


Figure 8: Selected cumulative regression coefficients for prevalence model, with ±2 robust standard errors.

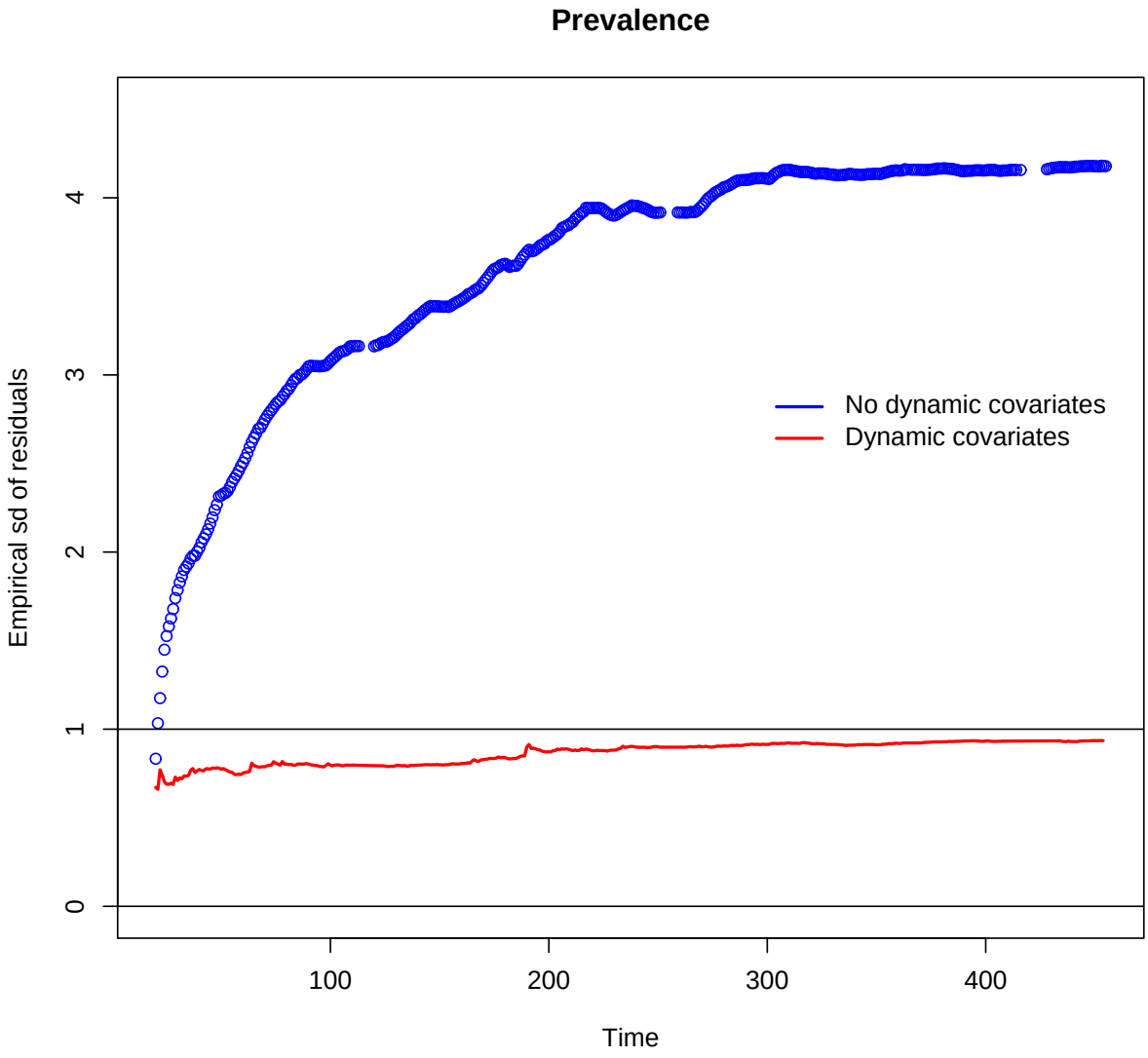


Figure 9: Prevalence model: empirical standard deviations of standardised martingale residual processes.

Table 1: Fixed covariate summary.

Description	Mnemonic	Summary
Male	Male	47%
Starting age (months)	Young	≤ 12 28%
		$12 - 24$ 36%
	Older	> 24 36%
3 or more people/bedroom	Dens	19%
Poor street quality	Stbad	57%
Contaminated water storage	Contstr	24%
Contaminated water source	Contsrce	22%
Standing water	Standwt	32%
Open sewerage	Opensew	16%
Rain affected accommodation	Rainacc	29%
Mother < 25 years	Youngmth	46%
Low socio-economic status	Poor	61%
Other children ≤ 5 years	Othchld	45%

Table 2: Estimates and standard errors for transition model.

	Events model	Missing data model		
		$m = 1$ and $l = 0$	$m = 2$ and $l = 0$	$m = 0$ and $l = 1$
Covariates	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Male	0.08 (0.05)	0.04 (0.03)	0.02 (0.19)	0.02 (0.03)
Young	0.24 (0.07)	-0.15 (0.05)	0.79 (0.36)	-0.02 (0.05)
Older	-0.56 (0.13)	0.03 (0.03)	0.02 (0.26)	-0.01 (0.03)
Dens	0.26 (0.08)	-0.02 (0.04)	0.09 (0.25)	-0.01 (0.04)
Stbad	-0.08 (0.05)	0.04 (0.03)	0.002 (0.19)	0.03 (0.03)
Contstr	-0.05 (0.06)	0.17 (0.03)	0.08 (0.20)	0.05 (0.03)
Contsrce	0.11 (0.06)	-0.02 (0.03)	-0.04 (0.23)	0.02 (0.03)
Standwt	0.01 (0.07)	0.01 (0.04)	-0.41 (0.30)	-0.002 (0.04)
Opensew	0.37 (0.10)	0.10 (0.05)	0.36 (0.26)	-0.02 (0.04)
Rainacc	0.14 (0.06)	-0.05 (0.04)	-0.12 (0.20)	0.07 (0.03)
Youngmth	0.17 (0.06)	-0.04 (0.03)	0.30 (0.18)	0.002 (0.03)
Poor	-0.002 (0.05)	-0.27 (0.04)	-0.48 (0.17)	-0.01 (0.03)
Othchld	-0.02 (0.04)	-0.07 (0.03)	-0.17 (0.16)	0.04 (0.03)
Period $150 - 300$ days	-0.14 (0.04)	0.02 (0.03)	0.88 (0.24)	0.18 (0.03)
Period > 300 days	-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)

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

	Dynamic model		
Covariates	Dropout	Incidence	Prevalence
Male	0.12	2.83	7.82
Young	1.48	2.83	15.67
Older	2.56	-14.66	-35.09
Dens	1.16	4.10	17.08
Stbad	0.12		-8.01
Contstr	1.00		-4.47
Contsrce	-0.08	1.99	7.41
Standwt	-1.63		2.32
Opensew	2.05	5.99	19.82
Rainacc	-2.06	3.78	10.32
Youngmth	1.37	3.75	14.31
Poor	-3.59		
Othchld	-1.34		

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

Diarrhoea previous days				Prevalence	
Four	Three	Two	One	Observed	Model
				2%	2%
			✓	58%	51%
		✓	✓	64%	60%
	✓	✓	✓	66%	63%
✓	✓	✓	✓	72%	66%