Published in final edited form as:
Stat Med. 2008 December 10; 27(28): 5890–5906. doi:10.1002/sim.3400.

Regression Splines in the Time-Dependent Coefficient Rates Model for Recurrent Event Data

Leila D. Amorim[1,*], **Jianwen Cai**[2], **Donglin Zeng**[2], and **Maurício L. Barreto**[3]

[1] Department of Statistics, Federal University of Bahia, Salvador, Brazil

[2] Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A

[3] Instituto de Saúde Coletiva, Federal University of Bahia, Salvador, Brazil

SUMMARY

Many epidemiologic studies involve the occurrence of recurrent events and much attention has been given for the development of modelling techniques that take into account the dependence structure of multiple event data. This paper presents a time-dependent coefficient rates model that incorporates regression splines in its estimation procedure. Such method would be appropriate in situations where the effect of an exposure or covariates changes over time in recurrent event data settings. The finite sample properties of the estimators are studied via simulation. Using data from a randomized community trial that was designed to evaluate the effect of vitamin A supplementation on recurrent diarrheal episodes in small children, we model the functional form of the treatment effect on the time to the occurrence of diarrhea. The results describe how this effect varies over time. In summary, we observed a major impact of the vitamin A supplementation on diarrhea after 2 months of the dosage, with the effect diminishing after the third dosage. The proposed method can be viewed as a flexible alternative to the marginal rates model with constant effect in situations where the effect of interest may vary over time.

1. INTRODUCTION

The most commonly used model in survival analysis is the Cox's proportional hazards model [1], which provides estimates of the relative risk associated with time-to-event occurrence. This method, however, is applied to those longitudinal studies in which the outcome can occur only once, for example, death or diagnosis of diabetes. The increasing complexity of the research conducted in many areas, considering intricate sampling schemes, capturing multiple occurrences of an outcome in the same subject or evaluating complex causal relationships, for instance, has yielded very complicated data structures which require sophisticated statistical methods to appropriately answer the research questions. Many studies involve the occurrence of recurrent events, such as times to opportunistic infections among AIDS patients or to lung exacerbations in cystic fibrosis patients, which has motivated methodological developments in survival analysis to handle complex recurrent event settings, including large number of recurrent events, presence of time-dependent covariates and time-dependent effects as well as potential dependent censoring, among other features [2,3,4,5,6,7,8,9,10]. Much effort has also been devoted to the development of methods for the estimation of means/rates of recurrent events in recent years [5,6,11]. The main appeal of using such approaches is that the mean number of events

[*]Correspondence to: Leila D. Amorim, UFBA, Instituto de Matemática, Rua Barão de Jeremoabo, s/n, Campus de Ondina, Salvador, Bahia, 40170-110, Brazil (leiladen@ufba.br).

and the average rate of event occurrence are usually of direct interest to investigators and are also easy to be understood, especially for non-statisticians.

In this context it is often of interest to explore the functional form of the relationships between covariates and time-to-event and to examine whether and how the effects are changing over time. Existing methods [12,13,14,15,16,17,18,19,20] are usually Cox-based models defined for univariate time-to-event. In such case, a sieve estimation procedure, assuming that the coefficient functions are piecewise constants [16], and a dynamic linear model approach, assuming that the baseline hazard and the coefficient functions are both piecewise constant functions, had been proposed [17]. More recently a local likelihood technique to estimate the time-dependent coefficients in Cox's regression model was developed [18]. Alternatively, several investigators [12,13,14,15] have used spline functions, which are well known for their usefulness in providing a smooth approximation to a covariate function, to model the relative risk in the Cox proportional hazards model. To our knowledge, however, no time-varying coefficient model had been proposed to handle recurrent time-to-event outcomes. Thus, in this paper we propose a method for estimating time-varying coefficients in the rates model using regression B-splines.

We illustrate the applicability of the proposed method using data from a randomized community trial that was designed to evaluate the effect of vitamin A supplementation on recurrent diarrheal episodes in 1240 pre-school age children, who were assigned to receive either placebo or vitamin A every 4 months for one year [21]. This study provides valuable information to evaluate multiple dosage of vitamin A and their effect on the incidence of diarrheal episodes. A log linear model with Poisson error, which is the standard regression model for incidence density rates, was used for analyzing this data and suggested that the overall incidence of diarrhea was significantly lower in the supplemented group than in the placebo group [21]. However, this method will not be the choice when the research question lies on important covariate or effects that change over time. In this case, as pointed out by Moulton and Dibley [22], use of time-to-event models will lead to greater efficiency and accuracy. Other statistical methods have also been applied for the analysis for vitamin A data [23]. Nevertheless, none of them had focused on the estimation of potential time-varying effect of the supplementation.

Rates models have been used to analyze multiple time-to-event data, where the rate of recurrence is modeled as a function of observed covariates and the effect of the covariates is assumed to be constant [6,24]. More recent analysis of the data from the vitamin A study using piecewise marginal rates model suggested that the effect of vitamin A supplementation on recurrent diarrhea may change over time. It is important to develop methods to improve the estimation of such time-varying effects. Hence, the main purpose of this paper is to present a statistical method that allows the estimation of time-varying coefficients in modeling recurrent time-to-event data. We consider the use of regression splines based on time-varying coefficient rates model to examine changes in effects over time. We also present results from a simulation study to evaluate the proposed method under different assumptions about the shape of the rate ratio function.

2. REGRESSION SPLINES IN THE TIME-DEPENDENT COEFFICIENT RATES MODEL

2.1. Data

Details of the vitamin A trial has been presented elsewhere [21]. Briefly, a randomized community trial was conducted in a cohort of 1,240 children, aged 6–48 months at baseline, who were assigned to receive either vitamin A or placebo every 4 months for 1 year in a small city in the Northeast of Brazil between December 1990 and December 1991. The

vitamin A dosage was 100,000 IU for children younger than 12 months and 200,000 IU for older children, which is the high dosage guideline established by the World Health Organization (WHO) for the prevention of vitamin A deficiency. The morbidity data was collected during household visits, which occurred three times per week, by local field workers during one year. The information on the occurrence of diarrhea and respiratory infections collected at each visit corresponds to a recall period of 48–72 hours. The number of liquid and semi-liquid motions per 24 hours was recorded. Besides the child's information, such as age and gender, there are also available socio-economic indicators for the households, which include mother's education, their working status, number of people living in the household, energy and water supply, among others, collected at the trial baseline. The study was approved by the National Institute of Nutrition, Ministry of Health, Brazil, and by the ethics committee of the School of Medicine, Federal University of Bahia.

For the analysis presented here, **a day with diarrhea** was defined when 3 or more liquid or semi-liquid motions were reported in a 24 hour period. **An episode of diarrhea** was defined as a sequence of days with diarrhea. An episode was considered finished when there were 3 or more days without diarrhea [21]. The **severity of a diarrheal episode** was defined based on the duration of an episode and on the number of liquid or semi-liquid motions reported in a 24 hour period. Thus, we defined three groups of episodes: mild (duration ≤2), moderate (duration ≥3 and average number of motions < 5) and severe (duration ≥3 and average number of motions ≥5) episodes.

We are defining the total time (i.e. time-to-event) as the time from the first dose of vitamin A until the occurrence of an episode of diarrhea. Since while a child is experiencing an episode, he/she is not at risk to another episode, we are considering the occurrence of discontinuous intervals of risk in the analysis. If an episode occurred, its last day was determined and the next risk interval for that child begun the next day. Furthermore, each child's observations are censored at the earliest of time of lost-to-follow up and end of study. The covariates considered in the models are treatment, which is 0 if the child received placebo and 1 if the child received vitamin A, gender, which is 0 if a girl and 1 if a boy, and age at baseline. We also conducted analyses defining the outcome to be the time from the first dose of vitamin A until the occurrence of a severe episode of diarrhea. Data management and preliminary data analysis were carried out using SAS version 8.2 software for Windows [25]. The proposed method was implemented using R version 1.9.1. software [26].

2.2. Model Estimation

We are focusing on a time-to-event approach for recurrent data that allow us to estimate effects that may change over time. In this way we are properly modeling the functional form of exposure or covariates by using a rates model that incorporates a smoothing technique called regression splines. The rates function for the *i*th individual is modeled as:

$$d\mu_i(t)=\exp\{\beta' \mathbf{Z}_i(t)+\theta(t)W_i(t)\}d\mu_0(t)$$

where β is a $(p-1)\times 1$ vector of fixed regression parameters, $\theta(t)$ is the time-varying regression parameter and $d\mu_0(t)$ is the baseline rate function. The covariates $\mathbf{Z}(t)$ and $W(t)$ could be time-independent or time-dependent. For instance, when W is a time-independent binary exposure or covariate, such as treatment group, the rate ratio(RR) of the two groups at time t is given by exponentiating $\theta(t)$ (i.e., $RR(t) = \exp(\theta(t))$). The estimation of $\theta(t)$ might be done by approximating $\theta(t)$ through standard cubic B-spline basis functions $B_k(t)$,(k = 1, …, m + 3), such that $\theta(t)=\gamma_0+\sum_{k=1}^{m+3}\gamma_k\tilde{B}_k(t)$, where m defines the number of interior knots.

Splines are piecewise polynomials satisfying continuity constraints at the knots joining the pieces. B-splines, originally introduced by de Boor [27], are a popular type of regression splines in statistical applications, mainly due to their flexibility and numerical properties. The proposed model uses products of a covariate and B-spline functions of time to yield models that allow effects to change over time in a flexible way.

Thus, replacing $\theta(t)$ by its B-spline approximation in the above time-varying coefficient model, we have:

$$d\mu(t)=\exp\{\beta' \mathbf{Z}(t)+\gamma' \tilde{\mathbf{W}}(t)\}d\mu_0(t),$$

where $\tilde{\mathbf{W}}_i(t)=(W_i(t), \tilde{B}_1(t)W_i(t), \ldots, \tilde{B}_{m+3}(t)W_i(t))'$ and $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{m+3})'$. Note that this model does not include time-dependent coefficients. Now the model becomes a standard marginal rates model with time-dependent covariate $\tilde{\mathbf{W}}(t)$. Thus, the estimates of the regression parameters are obtained by maximizing the following log pseudo-partial likelihood:

$$\ell(\beta,\gamma)=\sum_{i=1}^{n}\int_0^{\tau}\{\beta' \mathbf{Z}(t)+\gamma' \tilde{\mathbf{W}}(t) - log[nS^{(0)}(\beta,\gamma,t)]\}dN_i(t),$$

where $S^{(0)}(\beta,\gamma,t)=\frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp\{\beta' \mathbf{Z}(t)+\gamma' \tilde{\mathbf{W}}(t)\}$, $Y_i(t) = I(C_i \geq t)$ is the at-risk indicator, $dN_i(t)$ denotes the number of events in a small time interval $[t, t + dt)$, and τ is the study duration.

Considering the regularity conditions for the marginal rates model [6], the estimates of the regression parameters are obtained by the solution of the following unbiased estimating equation for $\eta = (\beta, \gamma)'$:

$$\mathbf{U}_n(\eta)=\frac{\partial \ell(\eta)}{\partial \eta}=\sum_{i=1}^{n}\int_0^{\tau}\left[\tilde{\mathbf{Z}}_i(u) - \frac{\mathbf{S}^{(1)}(\eta,u)}{S^{(0)}(\eta,u)}\right]dN_i(u),$$

where $\tilde{\mathbf{Z}}_i(t) = (\mathbf{Z}'_i(t), \tilde{\mathbf{W}}'_i(t))'$, $S^{(0)}(\eta,t)=n^{-1}\sum_{j=1}^{n} Y_j(t)\exp\{\eta' \tilde{\mathbf{Z}}_j(t)\}$, and $\mathbf{S}^{(1)}(\eta,t)=n^{-1}\sum_{j=1}^{n} Y_j(t)\exp\{\eta' \tilde{\mathbf{Z}}_j(t)\}\tilde{\mathbf{Z}}_j(t)$.

Because of the correlation between the recurrent events from the same subject, the negative inverse second derivative of the log pseudo likelihood is not a suitable variance estimator. Thus, we considered a robust covariance matrix estimator for $\hat{\eta}$ as defined by [6] as

$\widehat{\Gamma}=\frac{1}{n}\mathcal{I}(\widehat{\eta})^{-1}\widehat{\sum}\mathcal{I}(\widehat{\eta})^{-1}$, where $\mathcal{I}(\eta)= -\frac{1}{n}\frac{\partial^2 \ell(\eta)}{\partial\eta\partial\eta'}$ and
$\widehat{\sum}=\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{\tau}\left\{\tilde{\mathbf{Z}}_i(u) - \frac{\mathbf{S}^{(1)}(\eta,u)}{S^{(0)}(\eta,u)}\right\}d\widehat{M}_i(u)\right]^{\otimes 2}$, $d\widehat{M}_i(t)=dN_i(t)-\int_0^t Y_i(s)\exp\{\eta' \tilde{\mathbf{Z}}_i(s)\}d\widehat{\mu}_0(s)$ and the baseline mean function is estimated by the Breslow-type estimator as
$\widehat{\mu}_0(t)=n^{-1}\int_0^t dN.(u)/S^{(0)}(\eta,u)$, where $dN.(u)=\sum_{i=1}^{n} dN_i(u)$.

Therefore, $\theta(t)$ can be estimated by $\widehat{\theta}(t)=\widehat{\gamma}_0+\sum_{k=1}^{m+3}\widehat{\gamma}_k\tilde{B}_k(t)$ and its variance estimator is given by $var\{\hat{\theta}(t)\} = \tilde{\mathbf{B}}^*(t)' cov(\hat{\gamma})\tilde{\mathbf{B}}^*(t)$, where $\tilde{\mathbf{B}}^*(t) = (1, \tilde{B}_1(t), \ldots, \tilde{B}_{m+3}(t))'$ and $cov(\hat{\gamma})$ is the $(m+4)\times(m+4)$ matrix on the right bottom side of $\hat{\Gamma}$, assuming a fixed knot sequence. The

pointwise confidence intervals for $\theta(t)$ and related hypotheses tests are based on large-sample theory of maximum pseudo-partial likelihood estimation [3] and modern empirical process theory considered by Lin and colleagues [6]. The theory holds if the B-spline basis is chosen a priori.

Considering that the number of knots is held fixed as the sample size $n \to \infty$, we define a Wald-type statistic to test whether $\theta(t)$ is constant over time. Let $\underset{\sim}{\gamma} = (\gamma_1, \ldots, \gamma_{m+3})'$. Thus, the hypothesis of interest is that H_o: $\gamma = \mathbf{0}$. By analogy with the usual parametric likelihood procedures, this statistic can be defined by:

$$Q_W = \widehat{\underset{\sim}{\gamma}}' \, cov(\widehat{\underset{\sim}{\gamma}})^{-1} \widehat{\underset{\sim}{\gamma}}$$

The test rejects for large values of the statistic. Under the fixed knot framework, it is further assumed that the usual conditions are satisfied so that the standard asymptotic results hold for this model. Hence, under the null hypothesis, the statistic Q_W follows asymptotically a chi-square distribution with (m+3) degrees of freedom.

An advantage of such approach is that the estimates can be obtained by any statistical package that implements rates models and allows the inclusion of B-splines in the model.

2.3. Selection of Number and Location of Knots

For the developments presented here, we consider primarily splines with a fixed small number of knots as widely used in the literature [28,29,30]. The location of the interior knots is based on the quantiles of the observed event times in order to ensure an approximate equal number of events in each interval [30,20]. In practice, even though the choice of location of knots is dependent on the data, it is executed prior to fitting the model.

An alternative approach is to consider a posteriori model selection criteria, which may be used to find a reasonable trade-off between model parsimony and the risk of overfitting bias [13,29]. However, in such cases, additional variance is expected due to the posteriori model selection, which may inflate type I error rates. Among the criteria that were proposed for choosing the proper number of knots are the generalized cross-validation (GCV) and Akaike's information criterion (AIC)[15,28,31]. We considered here the GCV criterion to the recurrent time-to-event setting, which is an extension of the algorithm proposed by Nan and colleagues [15] for the univariate time-to-event data. We used the same form of GCV function assuming a working independence correlation matrix for recurrent events. However, if the time correlation matrix can be incorporated in constructing GCV as done for longitudinal data in Wang [32] and Fu[33], knots selection could be more accurate. An alternative criterion to define the optimal number of knots is the Akaike's information criteria (AIC), for which we specify several values of interior knots and chose the m that minimizes $AIC(m) = -2l(\beta, \gamma) + 2(m + degree + 1)$, with degree=3 for cubic splines.

3. SIMULATION STUDIES

3.1. Details of Data Generation

The proposed method was evaluated in simulation studies involving some variation of sample size, model complexity and shape of the true rate function. For each simulated data set, we estimated the time-varying coefficient $\theta(t)$ under the following marginal rates model:

$$E[dN(t)|Z] = d\mu(t) = \exp\{\theta(t)Z\} d\mu_0(t),$$

We generated recurrent event times using the random-effect intensity model

$$E[dN(t)|Z,u]=\lambda(t|Z,u)=u\lambda_0(t)\exp(\theta(t)Z),$$

where u is an unobserved unit-mean positive random variable that is independent of Z. We generated independent Z from Bernoulli distribution (0.5). We generated independent u_i (i=1,…,n) from gamma distribution, with mean 1 and variance $\sigma^2 = 0$ or 1. Since u is independent of Z and E(u)=1, then the random-effect intensity model implies the marginal rates model with $d\mu_0(t) = \lambda_0(t)dt$. The failure indicator Δ_{ij} was defined as $\Delta_{ij} = I(T_{ij} \leq C_i)$.

The recurrent event times were generated considering the relationships between $\lambda(t|Z, u)$, $\Lambda(t|Z, u)$ and $S(t|Z, u)$, denoting respectively the intensity function, the cumulative intensity function and the survival function, such that:

$$\Lambda(t|Z,u)=\int_0^t\lambda(s|Z,u)ds=\int_0^t u\lambda_0(s)\exp\{\theta(s)Z\}ds$$

and

$$S(t|Z,u)=\exp\{-\Lambda(t)|Z,u\}=\exp\{-\int_0^t u\lambda_0(s)\exp\{\theta(s)Z\}ds\}.$$

Since $S_{T_j|T_{j-1},T_{j-2},\ldots T_1}(t|Z,u)=\exp\{-\int_{T_{j-1}}^t u\lambda_0(s)\exp\{\theta(s)Z\}ds\}$, the jth recurrent event time can be generated by solving

$$\exp\{-\int_{T_{j-1}}^t u\lambda_0(s)\exp\{\theta(s)Z\}ds\}=\zeta_j$$

where ζ_j is generated from Unif(0,1), j=1,2,…J_i, $T_0 = 0$ and $t > T_{j-1}$.

The subject's follow-up time was uniform[0,1], such that it yielded an average of approximately 3.5–5.5 events observed per subject during the trial period. For data generation, the true log of the rate ratio of the two groups as functions of time was defined as −1.2 (a constant rate ratio over time), $log(t + 1)$ and $1.2sin(-\pi \times t)$. We refer to these models as constant, increasing and cycling. The cycling model reproduces the behavior of the effect of vitamin A supplementation on the occurrence of diarrhea episodes during the first dosage cycle.

Three different models for $\theta(t)$ were considered. The first is the cubic B-spline model, where

$$\theta(t)=\gamma_0+\sum_{k=1}^{m+3}\gamma_k\tilde{B}_k(t).$$

The second model specifies

$$\theta(t)=\overline{\gamma}_0+\sum_{k=1}^{m+2}\overline{\gamma}_k\tilde{A}_k(t),$$

where $\tilde{A}_1\ldots\tilde{A}_{m+2}$ is a quadratic B-spline basis. The third model specifies

$$\theta(t)=\dot{\gamma}_0+\sum_{k=1}^{m+1}\dot{\gamma}_k\tilde{C}_k(t),$$

where $\tilde{C}_1\ldots\tilde{C}_{m+1}$ is a linear/piecewise B-spline basis.

The estimation for $\theta(t)$ was performed considering the B-spline models above with 2 interior knots. For $\theta(t) = log(t+1)$, however, we compared the B-spline models for a range of different number of knots (m=2,…,6) through AIC criterion. The location of the interior knots was chosen to ensure an approximately equal number of failures between the knots [14,29].

For each combination defined by the model complexity and shape of the true rate function, 1,000 samples of size 100 or 200 were generated. For each configuration, we present the sampling bias, sampling/empirical standard error (ESE) of the estimates of $\theta(t)$, mean of the standard error estimator (SEE) and the coverage probability (CP) of the Wald 95% confidence interval.

In simulations that test the hypothesis that the time-dependent effect $\theta(t)$ is constant over time, $\tilde{\gamma} = \mathbf{0}$, empirical sizes of the spline based test considered 2,000 samples of sizes 100, 200 and 300 with different number of knots and spline models. The simulation studies were implemented using R version 1.9.1 software [26].

3.2. Results of Simulation Studies

We compared the estimates from the proposed method with those based on a standard marginal rates model to illustrate the importance of taking into account the time-dependent effect on a model when it is present. Figure 1 presents the true log of the rate ratio functions along with the mean of 1,000 estimates for the three shapes of the rate function discussed in this paper with $\sigma^2 = 1$. Note that the results from the proposed model describe the effects over time reasonably well for all situations. On the other hand, the standard marginal rates model with constant covariate effect will poorly describe the effects for the increasing and cycling models. Note in Figure 1 (a) with the true constant effect that the estimate from the standard marginal rates model is so close to the true value of the effect that we are not able to distinguish them. Some departure of the proposed method and the true line was noticed, especially, at the end of the study period, which is probably due to the very few events observed during this period.

In Table I, the results for piecewise/linear, quadratic and cubic B-splines models for $\theta(t) = log(1 + t)$ with sample sizes 100 and 200 are summarized. Generally, results indicate improved performance for sample size 200, for which the coverage probabilities, CP, closely approximated the nominal level, 0.95. The variance estimator performs well since the mean of the standard error estimates (SEE) and the empirical standard error of the estimates of $\theta(t)$ (ESE) are quite similar.

We compared different number of knots for the B-spline models with $\theta(t) = log(t + 1)$ in samples of size 100. In Table II the mean AIC values for 1,000 samples for a range of

number of interior knots (m=2,…,6) are presented. In all cases small number of knots seems to be most appropriate. According to the AIC criterion, two interior knots should be selected when considering quadratic and cubic B-splines models for this shape of the rate ratio while three knots would be the choice when considering a piecewise/linear model for this setup.

The results for the simulation studies considering the aforementioned B-splines models for $\theta(t) = 1.2sin(-\pi t)$ are displayed in Table III. The estimator of $\theta(t)$ presents small bias under quadratic and cubic B-splines models. The cubic B-spline model is the model with the smallest AIC when compared to piecewise/linear and quadratic B-spline models with 2 interior knots. The robust variance estimator provides a good estimation of the true variance of $\hat{\theta}(t)$, and the corresponding confidence intervals have reasonable coverage probabilities for quadratic and cubic B-spline models.

Table IV displays simulated empirical sizes for the Wald test for the hypothesis $\tilde{\gamma} = \mathbf{0}$. The Wald test shows substantial difference from the nominal level for the configurations that combine larger number of knots and smaller sample sizes. Improved results were obtained as sample size increases for all spline models. Overall, the empirical sizes were closer to the nominal level when considering small number of knots (m=2) and large sample sizes (n=300).

For the models with 2 interior knots, results given in Table IV indicate close agreement with the nominal level for the test with regression splines from n=200, particularly using the linear spline model. The models with 5 interior knots had sizes larger than the nominal level. It is possible that larger sample sizes are needed for the asymptotic distribution to be an accurate approximation in models with larger number of knots. The linear spline model presents the best results in terms of empirical sizes for this setup.

4. APPLICATION: VITAMIN A SUPPLEMENTATION AND DIARRHEA IN CHILDREN

The analyses include 1,207 children with mean age 27.3 months at baseline (std dev=12.1 months), 52.4 % boys and 50.1% randomized to receive vitamin A supplementation. Among those children, 1,063 (88.1%) had at least one diarrheal episode during their follow-up period in the study. The average number of days of follow-up is 331 days, with 83.7% of the children having daily continuous information for a year. The mean number of diarrheal episodes per child during the follow-up period is 5.9 (std dev=5.4, range=0–27 episodes). The median number of episodes in the vitamin A and placebo group is 4.0 and 5.0, respectively. According to the distribution of the number of episodes by treatment group, 16.45% of the children in the placebo group had 12 or more episodes of diarrhea during their follow-up period while this proportion was 14.55% in the vitamin A group.

The overall number of episodes of diarrhea to be consider in this study is 7,109 episodes, being 3,464 and 3,645 episodes, respectively, in the vitamin A and placebo groups. We classified the episodes according to their severity based on the duration of an episode and on the number of liquid or semi-liquid motions reported in a 24 h period. Using such criteria, we verified that only 276 (3.88 %) of the episodes were considered severe. The number of episodes decreased significantly during the study, such that 42.86% (3047) of the episodes occurred during the first dosage cycle (i.e, between first and second doses); 37.00% (2630) during the second dosage cycle and only 20.14% (1432) occurred between third and fourth doses of vitamin A (i.e, third dosage cycle).

For the evaluation of the overall effect of vitamin A supplementation on diarrhea, we first applied the standard marginal rates model, without time-dependent effects, to this data. The

result shows that the occurrence rate of episodes of diarrhea since first dosage is 8.8% lower for those who received vitamin A compared to those who received placebo, after adjusting by gender and age at baseline. This overall effect was borderline statistically significant ($\hat{\theta} = -0.092$; 95% CI=(−0.191; 0.006)). As expected, there is a negative effect of age on the rate of event occurrence, i.e., the rate of experiencing an episode of diarrhea decreases as the child becomes older. In order to evaluate and describe how the effect of vitamin A supplementation behaves over time, we implemented the proposed model. Figure 2 contains the curve for the log of the rate ratio of the occurrence of diarrhea smoothed over time considering 6 interior knots. The number of knots was selected using AIC criteria. The result suggests that, after the first dosage of vitamin A, there is an important reduction on the risk of diarrhea for the supplemented children. However, this effect disappears by the end of the first 4-month treatment cycle. After the second dose, an even more intense reduction on the risk of diarrhea was observed. At the end of the second treatment cycle, the effect of vitamin A supplementation reduces substantially and perhaps reverses. The reduced number of episodes during the third cycle may have affected the estimates after third dose. Results from Wald test suggest a significant time-dependent effect of treatment on the occurrence of diarrhea (p=0.0118).

We also fitted the proposed model for distinct age groups since age is an important known factor in the reduction of diarrhea incidence in children. For the analysis presented here we considered the three following age groups: (i) children with age at baseline being 12 or less months, (ii) children with age at baseline between 12 and 24 (inclusive) months and (iii) children older than 24 months at baseline. Note from Figure 3 that the effect of the supplementation behaves somewhat differently among the three age groups. According to Figure 3, the treatment effect seems to be slightly greater for the younger children (age ≤12 and 12 < age ≤ 24 months) when compared to older children (age > 24 months), especially regarding the first dosage. For all age groups we considered models with 3 interior knots based on AIC criteria. Gender was not a significant factor on any of the models considered.

We computed AIC and GCV as a posteriori model selection criteria to define the number of knots to be considered in our models, considering a range of values for interior knots (m=2, …,6). We verified that both criteria indicated a very similar number of knots for our data (results not shown here). We also compared the estimated trajectories of treatment effect for rates models using different number of knots and it was observed that even for a model with a large number of knots (m=6), the behavior of the treatment effect over time do not vary drastically from the models with smaller number of knots (m=2).

Lastly we fitted models considering the time until the occurrence of severe episodes as the outcome. In that case, the overall effect of vitamin A supplementation was larger than that obtained when considering the occurrence of any episode of diarrhea. The results from fitting a marginal rates model, without time-dependent effect, pointed out for a reduction of 31.8% on the occurrence rate of severe episodes of diarrhea for those who received vitamin A compared to those who received placebo, after adjusting by gender and age at baseline. This overall effect was statistically significant ($\hat{\beta} = -0.388$; 95% CI=(−0.703; −0.073)). Age at baseline had a significant negative effect on time to the occurrence of severe episodes of diarrhea while gender was not a significant factor again. The rate ratio for the occurrence of severe episodes of diarrhea for children with 12 months when compared to children with 48 months at baseline was 5.4. Figure 4 shows the estimated log rate ratio function for treatment effect on severe episodes of diarrhea through the use of rates models with regression cubic B-splines considering 6 interior knots. The reduction on the rate of severe episodes of diarrhea seems to happen earlier than that observed for any episode after the supplementation of the first dosage of vitamin A. The treatment effect also seems to be subject to more variability for severe episodes than that observed for all episodes, which

could be the consequence of the small number of such events. There were 276 severe episodes of diarrhea over the trial period, which occurred in only 15.24% of children in the study. As opposed to the results for the model for all episodes, the Wald test points out for the lack of evidence of a time-varying effect of treatment on the occurrence of severe episodes (p=0.2782). Again this result may have been affected by the reduced number of recurrent severe episodes of diarrhea.

5. DISCUSSION

Several investigators [12,13,14,15,29,30] have used spline functions to model the relative risk in the proportional hazards model. Such approaches provide greater flexibility for fitting data without *a priori* assumption about the form of the variation of the hazard ratio over time [30]. All available methods in the literature, however, were defined for univariate time-to-event settings. The proposed approach is useful in estimating time-varying coefficients in the recurrent time-to-event data setting.

By introducing regression splines, which are splines with small number of knots, in the marginal rates model and extending the known methods for recurrent time-to-event data, we developed a method that allows the investigator to describe with details the behavior of the effects of interest over time on the rate of event occurrence. Results from simulations indicate that unless there are very few events, the estimates of the rate ratios are approximately unbiased and the variance estimator performs well. Our approach can be viewed as a flexible alternative to the marginal rates model in situations where the effect of interest may vary over time. The proposed method can be implemented using the standard survival library in R.

The splines are well known for their usefulness in providing a smooth approximation to a covariate function. A spline is a piecewise polynomial and its shape depends on the degree of the spline function, on the number and location of the breakpoints or knots. A cubic spline (i.e., degree=3) should in most cases be sufficient to reflect changes in the log hazard as a function of the covariate of interest [13]. However, as pointed out by many authors [13,14,30], the use of regression splines implies a judicious choice of the number and location of knots because the shape of some estimates can depend heavily on this selection. Even though an increase of the number of the knots may result in more flexibility of the spline function, it may overfit the data and cause loss of statistical power if the underlying relationship is relatively simple [30]. High variability in the estimates at the observation period boundaries is often expected with this type of approach because the event times are sparse at the end of the study. Knots placement does not totally control the variability. Consequently, the estimates at the boundaries should not be emphasized. Some criteria for model selection have been proposed in the context of standard Cox regression, which includes cross-validation (CV), generalized cross-validation (GCV) and Akaike information (AIC) criteria [15,31]. In this paper, we focused primarily on the use of splines with a fixed small number of knots. However, we also considered GCV and AIC as a posteriori model selection criteria. In large data sets, choosing the number of knots using the GCV method can require considerable computational resources and may even be not feasible in some situations.

The proposed method was applied to evaluate the relationship between high doses of vitamin A and occurrence of diarrhea episodes in small children using data from a randomized community trial conducted in Brazil [21]. The impact of vitamin A supplementation on mortality of children with age between 6 months and 5 years-old had been verified by numerous studies in the last two decades, leading to a consensus about the protective role of vitamin A supplementation on childhood mortality. In contrast to the clear

effect of vitamin A on mortality, controversial results regarding the impact of vitamin A supplementation on diarrhea incidence have been shown. Studies conducted in India [34,35], China [36], Bangladesh [37] and Brazil [21] showed some evidence of significant reduction in overall incidence of diarrhea. At the same time, other studies [38,39,40,41,42] did not find significant reductions in either the incidence or mean daily prevalence of diarrhea. However, there is considerable evidence of a significant impact of vitamin A supplementation on the reduction in the incidence of severe diarrhea. For instance, in one trial, there was a 36% reduction in the mean daily prevalence of diarrhea (associated with fever) among supplemented children older than 23 months [40]. In the Brazilian study, the mean daily prevalence rates were 20% lower in the vitamin A supplemented group when defining diarrhea by 5 or more liquid or semi-liquid motions reported in a 24 hour period [21]. In another study, there was a significant difference in the average duration of diarrhea per episode between the two groups [34].

For the analysis of the effect of vitamin A supplementation on diarrhea in the study conducted in Brazil [21], the implementation of the proposed method provided further evidence on the effectiveness of such policy to prevent diarrhea in young children and more detailed insights into the behavior of such effect over time. The curves of the rate ratio of the occurrence of diarrhea smoothed over time were very helpful in determining the potential duration of the effect for each of such dosages. The most suitable dosing interval for the prophylatic vitamin A that has been recommended is of 4 to 6 months, without a consensus [42]. Our results indicate that there was no cumulative effect of successive dosages of vitamin A supplementation. It also suggests that an interval of maximum of 4 months between high doses of vitamin A may be adequate in this population. As diarrhea is still a major cause of morbidity and mortality in small children in developing countries, these results might be useful to help in designing effective health policies in programs of vitamin A supplementation.

In summary, a time-dependent coefficient rates model with small number of knots was proposed to estimate effects in the marginal rates model that may vary over time. The inclusion of splines in the estimation procedure provided flexibility to capture the time-varying effect and allowed that the inferences were made using standard techniques. This methodology may be potentially useful for describing the behavior of many other exposures or covariates associated to research questions in Epidemiology and Public Health.

Acknowledgments

During this research Dr. Amorim was supported by a CAPES scholarship from Brazil. This work was supported in part by the National Institutes of Health grant R01-HL57444 and by PRONEX, CNPq, Brazilian Federal Government (Contract number 661086/1998-4).

Contract/grant sponsor: CAPES/Brazil; contract/grant number: BEX–1789/00-7

Contract/grant sponsor: National Institutes of Health; contract/grant number: R01–HL57444

References

1. Cox DR. Regression Models and Life-tables (with discussion). Journal of Royal Statistics Society - Series B. 1972; 34:182–220.
2. Prentice RL, Williams BJ, Peterson AV. On the Regression Analysis of Multivariate Failure Time Data. Biometrika. 1981; 68:373–379.
3. Andersen PK, Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics. 1982; 10:1100–1120.

4. Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association. 1989; 84:1065–1073.
5. Pepe MS, Cai J. Some Graphical Displays and Marginal Regression Analysis for Recurrent Failure Times and Time Dependent Covariates. Journal of American Statistical Association. 1993; 88:811–820.
6. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric Regression for the Mean and Rate Functions of Recurrent Events. Journal of Royal Statistics Society - Series B. 2000; 62:711–730.
7. Wang MC, Qin J, Chiang CT. Analyzing Recurrent Event Data with Informative Censoring. Journal of the American Statistical Association. 2001; 96:1057–1065.
8. Duchateau L, Janssen P, Kezic I, Fortpied C. Evolution of Recurrent Asthma Event Rate Over Time in Frailty Models. Applied Statistics. 2003; 52:355–363.
9. Ghosh D, Lin DY. Semiparametric Analysis of Recurrent Event Data in the Presence of Dependent Censoring. Biometrics. 2003; 59:877–885. [PubMed: 14969466]
10. Miloslavsky M, Keles S, van der Laan MJ, Butler S. Recurrent Events Analysis in the Presence of Time-dependent Covariates and Dependent Censoring. Journal of Royal Statistics Society - Series B. 2004; 66:239–257.
11. Lawless JF, Nadeau C. Some Simple Robust Methods for the Analysis of Recurrent Events. Technometrics. 1995; 37:158–168.
12. Hastie T, Tibshirani R. Varying-coefficient Models. Journal of Royal Statistics Society - Series B. 1993; 55:757–796.
13. Sleeper LA, Harrington DP. Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease. Journal of the American Statistical Association. 1990; 85:941–949.
14. Gray RJ. Flexible Methods for Analyzing Survival Data using Splines, with Applications to Breast Cancer Prognosis. Journal of the American Statistical Association. 1992; 87:942–951.
15. Nan B, Lin X, Lisabeth LD, Harlow SD. A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause. Biometrics. 2005; 61:576–583. [PubMed: 16011707]
16. Murphy SA, Sen PK. Time-dependent coefficients in a Cox-type regression model. Stochastic Processes and Applications. 1991; 39:153–180.
17. Gamerman D. Dynamic Bayesian methods for survival data. Applied Statistics. 1991; 40:63–79.
18. Cai Z, Sun Y. Local Linear Estimation for Time-Dependent Coefficients in Cox's Regression Models. Scandinavian Journal of Statistics. 2003; 30:93–111.
19. Verweij PJM, van Houwelingen HC. Time-Dependent Effects of Fixed Covariates in Cox Regression. Biometrics. 1995; 51:1550–1556. [PubMed: 8589239]
20. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Statistics in Medicine. 1994; 85:1045–1062. [PubMed: 8073200]
21. Barreto ML, Santos LMP, Assis AMO, Araujo MPN, Farenzena GG, Santos PAB, Fiaccone RL. Effect of Vitamin A Supplementation on Diarrhoea and Acute Lower-respiratory-tract Infections in Young Children in Brazil. Lancet. 1994; 344:228–231. [PubMed: 7913157]
22. Moulton LH, Dibley MJ. Multivariate Time-to-Event Models for Studies of Recurrent Childhood Diseases. International Journal of Epidemiology. 1997; 26:1334–1339. [PubMed: 9447414]
23. Andreozzi VL, Bailey TC, Nobre FF, Struchiner CJ, Barreto ML, Assis AMO, Santos LMP. Random-Effects Models in Investigating the Effect of Vitamin A in Childhood Diarrhea. Annals of Epidemiology. 2006; 16:241–247. [PubMed: 16303315]
24. Cai J, Schaubel D. Analysis of Recurrent Event Data. Handbook of Statistics. 2004; 23:603–623.
25. SAS Institute Inc. SAS/STAT software. Cary, NC: SAS Institute, Inc; 1999–2001.
26. R Development Core Team. R: a Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2004. http://www.R-project.org/
27. De Boor, C. AA Practical Guide to Splines. Revised. Springer; 2001.
28. Rosenberg PS. AHazard Function Estimation Using B-Splines. Biometrics. 1995; 51:874–887. [PubMed: 7548706]

29. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing with Application in Lupus Nephritis. The Journal of the American Statistical Association. 1996; 91:1432–1439.
30. Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, Faivre J. A Relative Survival Regression Model using B-spline Functions to Model Non-proportional Hazards. Statistics in Medicine. 2003; 22:2767–2784. [PubMed: 12939785]
31. O'Sullivan. Nonparametric Estimation of Relative Risk using Splines and Cross-validation. SIAM Journal on Scientific and Statistical Computing. 1988; 9:531–542.
32. Wang Y. Smoothing spline models with correlated random errors. Journal on the American Statistical Association. 1998; 93:341–348.
33. Fu WJ. Nonlinear GCV and quasi-GCV for shrinkage models. Journal of Statistical Planning and Inference. 2005; 131:333–347.
34. Biswas R, Biswas AB, Manna B, Bhattacharya SK, Dey R, Sarkar S. Effect of Vitamin A Supplementation on Diarrhoea and Acute Lower Respiratory Tract Infection in Children. European Journal of Epidemiology. 1994; 10:57–61. [PubMed: 7957792]
35. Chowdhury S, Kumar R, Ganguly NK, Kumar L, Walia BN. Effect of Vitamin A Supplementation on Childhood Morbidity and Mortality. Indian Journal of Medical Science. 2002; 56:259–264.
36. Lie C, Ying C, Wang EL, Brun T, Geissler C. Impact of Large-dose Vitamin A Supplementation on Childhood Diarrhoea, Respiratory Disease and Growth. European Journal of Clinical Nutrition. 1993; 47:88–96. [PubMed: 8436094]
37. Rahman MM, Vermund SH, Wahed MA, Fuchs GJ, Baqui AH, Alvarez JO. Simultaneous Zinc and Vitamin A Supplementation in Bangladeshi Children: Randomised Double Blind Controlled Trial. BMJ. 2001; 323:314–318. [PubMed: 11498488]
38. Abdeljaber MH, Monto AS, Tilden RL, Schork A, Tarwotjo I. The Impact of Vitamin A Supplementation on Morbidity: A Randomized Community Intervention Trial. American Journal of Public Health. 1991:1654–1656. [PubMed: 1746667]
39. Ramakrishnan U, Latham MC, Abel R, Frongillo EA Jr. Vitamin A Supplementation and Morbidity among Preschool Children in South India. American Journal of Clinical Nutrition. 1995; 61:1295–1303. [PubMed: 7762534]
40. Bhandari N, Bhan NK, Sazawal S. Impact of Massive Dose of Vitamin A given to Preschool Children with Acute Diarrhoea on Subsequent Respiratory and Diarrhoeal Morbidity. BMJ. 1994; 309:1404–1407. [PubMed: 7819847]
41. Dibley MJ, Sadjimin T, Kjolhede CL, Moulton LH. Vitamin A Supplementation Fails to Reduce Incidence of Acute Respiratory Illness and Diarrhea in Preschool-age Indonesian Children. Journal of Nutrition. 1996; 126:434–442. [PubMed: 8632216]
42. Ross DA, Kirkwood BR, Binka FN, Arthur P, Dollimore N, Morris SS, Shier RP, Gyapong JO, Smith PG. Child Morbidity and Mortality Following Vitamin A Supplementation in Ghana: Time since Dosing, Number of Doses, and Time of Year. American Journal of Public Health. 1995; 85:1246–1251. [PubMed: 7661232]