ESTIMATING TIME-VARYING TREATMENT EFFECT FOR RECURRENT CHILDHOOD DISEASES

by
Leila Denise Alves Ferreira Amorim

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, School of Public Health.

Chapel Hill
2006

Approved by:

Dr. Jianwen Cai, Advisor
Dr. Donglin Zeng, Advisor
Dr. Shrikant Bangdiwala, Committee Member
Dr. Amy Herring, Committee Member
Dr. James Thomas, Committee Member

ABSTRACT
LEILA DENISE ALVES FERREIRA AMORIM: ESTIMATING TIME-VARYING TREATMENT EFFECT FOR RECURRENT CHILDHOOD DISEASES.
(Under the direction of Dr. Jianwen Cai and Dr. Donglin Zeng.)

Many medical studies involve the occurrence of recurrent events, such as times to opportunistic infections among AIDS patients. In particular, this doctoral research has been motivated by the need for analyzing the effect of vitamin A supplementation on recurrent diarrheal episodes from a randomized community trial conducted in a cohort of 1,240 children, aged 6-48 months at baseline, in Brazil. Rate models have been used to analyze such type of 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. Preliminary analysis of the vitamin A study suggested that the effect of vitamin A supplementation on diarrhea may change over time. It is important to develop methods to estimate such time-varying effects. Hence, the main purpose of this research is to develop statistical methods that incorporate time-varying coefficients in modeling recurrent time-to-event data.

Rate models with time-varying coefficients are proposed to analyze recurrent time-to-event data. B-splines are used for the estimation of regression time-varying coefficients using two approaches: regression and penalized splines. Estimation of smoothing parameter, number and placement of knots is discussed. The small sample properties of the estimators are studied via simulation. Data from the vitamin A study is analyzed using the proposed methods.

Another focus of the dissertation research is on the comparison of statistical methods for recurrent event data with dependent or informative censoring. Many statistical

methods assume independent censoring. However, this assumption may not hold in some studies. Two methods have recently been proposed to account for dependent censoring for marginal rate models with recurrent event data. The first approach was developed by Wang, Qin and Chiang (2001), who proposed to model the occurrence of recurrent events by a subject-specific nonstationary Poisson process via a latent variable. In the second approach Miloslavsky, Keles, van der Laan and Butler (2004) proposed inverse probability of censoring weighted (IPCW) estimators for the regression parameters in the proportional rate model in order to obtain consistent estimators in the presence of dependent censoring. These two methods are critically compared through extensive simulation studies.

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my advisors Drs. Jianwen Cai and Donglin Zeng for their guida patience and encouragement, and the members of my committee Drs. Shrikant Bangdiwala, Amy Herring and James Thomas for their valuable comments and suggestions. In particular, I am grateful to my program advisor, Dr.Bangdiwala, for his guidance, friendship, constant support and incentive. Under his supervision, the experience I gained during the four years employed by Dr.Harrell as a graduate research assistant was invaluable. Special thanks go to Dr.Harrell, Dr.McMurray and colleagues from the CHIC Study in the Nursing School-UNC for providing me opportunity to contribute and learn about cardiovascular research in children in an atmosphere of scholarship and friendship.

I am also indebted to my colleagues from the Department of Statistics in the Federal University of Bahia, Brazil, for supporting me in my desire to advance my education. I also owe thanks to Mauricio Barreto e Nelson Oliveira for offering me many challenging projects which stimulated my interest in biostatistics. This dissertation would not be possible without financial support from the Federal University of Bahia and CAPES, both institutions from Brazil.

I thank my friends (so many to be listed here) and family, especially, my mother, father, sister and her family, for their encouragement, prayers and love. Finally, I want to dedicate this work to my husband Renato for his unconditional love, support and encouragement, and to my son Matheus for providing me such wonderful moments and the strength to overcome all the difficulties.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Many medical 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 in the last few decades (Prentice, Williams and Peterson, 1981; Andersen and Gill, 1982; Wei, Lin and Weissfeld,1989; Pepe and Cai, 1993; Lin et al., 2000). Particularly the marginal hazards models, such as WLW (Wei et al., 1989) and LWA (Lee, Wei and Amato, 1992), and conditional hazards models, such as PWP (Prentice et al., 1981) and AG (Andersen and Gill, 1982) have been used more frequently, especially for having been already incorporated by many statistical software packages.

Even though these methods are robust and well-developed, recent papers (Kelly and Lim, 2000) have been discussing the appropriateness of such approaches to handle recurrent event data. In general, it has been recommended that WLW is more appropriate in situations where there are different types of events from the same person while LWA is more suitable for clustered data. When used for the analysis of recurrent event data, both of them presented a *carry-over* effect for subsequent events, especially when the estimated effect for the first event was large. Wei and Glidden (1997) pointed out that AG and PWP are both sensitive to misspecification of the dependence structure among the recurrence times. In addition, AG assumes that the

event recurrences follow a non-homogeneous Poisson process, with the hazard being unaffected by earlier events that occurred to the subject given the covariates, although the covariates could include information from earlier events. Recent research has been focusing on more complex recurrent event settings which include large number of recurrent events, time-dependent covariates, time-dependent coefficients and dependent censoring among other features (Wang, Qin and Chiang, 2001; Duchateau et al., 2003; Ghosh and Lin, 2003; Miloslavsky et al., 2004). Particularly, Duchateau et al (2003) discussed the use of parametric and semiparametric frailty models for recurrent event data while the methods proposed by Wang, Qin and Chiang(2001), Ghosh and Lin (2003) and Miloslavsky et al(2004) focus on modelling the recurrent event data in the presence of dependent censoring.

Much effort has also been devoted to the development of methods for the estimation of means/rates of recurrent events in recent years. Pepe and Cai (1993) studied methods to display and estimate rate functions for the analysis of multiple time-to-event data. Later, Lawless and Nadeau (1995) presented robust methods for nonparametric estimation of the cumulative mean/rate function as well as proposed a marginal model for the mean/rate estimation. More recently, Lin et al (2000) provided rigorous justification for the use of semiparametric regression for the mean and rate functions in the analysis of recurrent time-to-event data. Based on that, Miloslavsky et al (2004) presented an estimator for the regression parameters in the marginal proportional rates model considering the presence of time-dependent covariates and dependent censoring.

This doctoral research has been motivated by the need for analyzing the recurrent diarrheal episodes in small children, with the main questions of interest being related to the effect of the supplementation of vitamin A from a randomized community trial with 1240 children aged 6 to 48 months at baseline and who were followed-up for 1 year (Barreto et al, 1994). This study provides valuable information to evaluate multiple

dosage of vitamin A and their effect on the incidence of diarrheal episodes. It is also a major interest to evaluate whether the effect of vitamin A supplementation persists over time and if so for how long it does. In this study the children were assigned to receive either placebo or vitamin A every 4 months for one year. The original analysis performed to evaluate the effect of such supplementation did not focus on how the supplementation effect behaves over time. However, recent analysis conducted to this data through the use of piecewise marginal rates models pointed out that the effect of vitamin A supplementation on the occurrence of diarrhea may change over time.

The plausibility of a time-dependent effect for treatment on the study about supplementation of vitamin A motivated our research that focus on time-varying coefficient models that incorporates B-splines to describe how the effects change over time. Existing methods (Hastie and Tibshirani, 1990; Sleeper and Harrington, 1990; Gray, 1992; Nan et al., 2003) are all Cox-based models defined for univariate failure time. Even though much recent research effort has been devoted to the analysis of multivariate failure time data, to our knowledge, no time-varying coefficient model has been proposed to multivariate time-to-event outcomes. Thus, in Chapters 3 and 4, we proposed two methods for estimating time-varying coefficients for recurrent time-to-event data considering rate models with B-splines.

The discussion in the literature about the potential occurrence of dependent or informative censoring in many medical studies, on the other hand, lead us to the implementation and comparison of two fairly recent proposed rates models to handle dependent censoring in recurrent event settings (Wang et al, 2001; Miloslavsky et al, 2004). The comparison of these methods were performed through extensive simulation studies, whose results are presented in Chapter 5. Since 16.3% of the children were not followed up during all study period in the vitamin A trial, we also applied the methods that take into account possible dependent censoring to this data.

In the next Chapter we summarized the current and relevant literature in those topics.

CHAPTER 2

LITERATURE REVIEW

2.1 Modeling Recurrent Event Data Assuming Independent Censoring

Let's first define the notation that will be used for the subsequent sections. Suppose that the data consist of n subjects. Let $X_{i,k} = T_{i,k} \wedge C_i$ be the observed time for the i^{th} subject with respect to the k^{th} event, where C_i denotes censoring time and $T_{i,1}, T_{i,2}, ..., T_{i,k}$ represents the event times. These event times are called total times and represent, for instance, the time since randomization to treatment until the occurrence of the k^{th} event for the i^{th} subject. Let's define the gaps between successive events as $G_{i,k} = T_{i,k} - T_{i,k-1}$ for $k = 1, 2, ...K$, with $T_{i,0} \equiv 0$. Let τ denote the end of the study and $\mathbf{Z}_i$ be a possibly time-dependent covariate vector. Let $\Delta_{ik} = I(T_{i,k} \leq C_i)$ be an event indicator, which takes value 1 if an event is observed and 0 otherwise. We are assuming independent censoring.

Using counting process notation, let $N_i(t) = \int_0^t dN_i(s)$ represents the number of events in $[0, t]$ for subject i, where $dN_i(s)$ denotes the number of events in the small time interval $[s, s + ds]$. Let's define the event intensity function as:

$$\lambda_i(t|H_i(s)) = \lim_{\Delta t \to 0} \frac{Pr(N_i(t + \Delta t) - N_i(t) = 1|H_i(t))}{\Delta t}$$

where $H_i(s) = (N_i(s), 0 \leq s < t; \mathbf{Z}_i(s), 0 \leq s \leq t)$ represents the process history up to time t. It is assumed that the probability of more than one event over the interval $[t, t + \Delta t)$ is $o(\Delta t)$, so $E[dN_i(t)|H_i(t)] = \lambda_i(t; H_i(t))dt$. In the next sections we will use such notation for defining the survival models to evaluate the effect of factors of interest in the occurrence of recurrent events.

2.1.1 Conditional Hazards Models

2.1.1.1. Andersen-Gill Model

The Cox proportional hazards model (Cox, 1972) has been the most popular procedure for modeling the relationship between covariates and the failure rate or hazard function. The Cox model assumes that the hazard of individual i at time t is given by:

$$\lambda_i(t) = \lambda_0(t)e^{\beta' \mathbf{Z}_i(t)}$$

where $\lambda_0(.)$ is an unspecified nonnegative function of time called baseline hazard and β is a $p \times 1$ column vector of unknown parameters. Estimation of β is based on the partial likelihood function introduced by Cox (1975), such that the log partial likelihood can be written, using counting process notation, as:

$$\ell(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left[\beta' \mathbf{Z}_i(t) - \log \left(\sum_j Y_j(t) e^{\beta' \mathbf{Z}_i(t)} \right) \right] dN_i(t)$$

where $Y_i(t) = I\{X_i \geq t\}$. The large-sample properties of parameter estimators can be obtained through the theory of Martingales (Andersen and Gill, 1982) or empirical processes (Tsiatis, 1981).

An extension of Cox proportional hazards model for multiple event data is the Andersen-Gill model (Andersen and Gill, 1982), in which the intensity function for the

kth recurrence relates to the covariates through the following formulation:

$$\lambda_{ik}(t|\mathbf{Z}_i(t)) = Y_{ik}(t)\lambda_0(t)e^{\beta'\mathbf{Z}_i(t)},$$

for $k = 1, ..., K_i$. This model assumes a common baseline hazard for all events and that the number of events in nonoverlapping time intervals are independent, given the covariates, which is known as *independent increments* assumption (i.e., non-homogeneous Poisson process (Chiang, 1968)). Although the subjects may experience more than one event, a subject can only make one contribution to the risk set for a given event at any specific time. Moreover, under this model, the risk sets for the *(k+1)*th recurrences are not restricted to the subjects who have experienced the first k recurrences. In such case, a subject's second event time may contribute to the risk set corresponding to another subject's first event, for instance (Kelly and Lim, 2000).

The parameter estimation is carried out by partial likelihood. An iterative algorithm can be used to obtain an estimator of β, denoted by $\hat{\beta}$, by solving the estimating equation $\mathbf{U}(\beta) = \mathbf{0}$, where:

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

with $\mathbf{S}^{(j)}(\beta,t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{ik}(t)\mathbf{Z}_i(t)^{\otimes j}e^{\beta'\mathbf{Z}_{ik}(t)}$, and for a vector $\mathbf{z}$, $\mathbf{z}^{\otimes 0} = 1$, $\mathbf{z}^{\otimes 1} = \mathbf{z}$, $\mathbf{z}^{\otimes 2} = \mathbf{z}\mathbf{z}'$.

The Breslow-Aalen estimate of the cumulative baseline hazard is given by $d\hat{\Lambda}_0(\beta,t) = n^{-1}\int_0^t dN_\cdot(t)/S^{(0)}(\beta,t)$, where $dN_\cdot(t) = \sum_{i=1}^{n} dN_i(t)$. The information matrix is defined as:

$$\mathcal{I}(\beta) = -\partial^2\ell(\beta)/\partial\beta\partial\beta' = \sum_{i=1}^{n}\int_0^\tau \left[\frac{\mathbf{S}^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - \left\{\frac{\mathbf{S}^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\right\}^{\otimes 2}\right] dN_i(t)$$

Under certain regularity conditions, as $n \to \infty$, $n^{-\frac{1}{2}}\mathbf{U}(\beta)$ has an asymptotic normal distribution with mean zero and a variance which can be consistently estimated by $n^{-1}\mathcal{I}(\beta)$ and $n^{\frac{1}{2}}(\hat{\beta} - \beta)$ has an asymptotic normal distribution with mean zero and a variance which can be consistently estimated by $n\mathcal{I}(\beta)^{-1}$ (Andersen and Gill, 1982).

A robust variance estimator for $\mathbf{U}(\beta)$ is given by $n\hat{\mathbf{\Sigma}}(\beta)$, where

$$\hat{\mathbf{\Sigma}}(\beta) = n^{-1} \sum_{i=1}^{n} \hat{\mathbf{B}}_\mathbf{i}(\beta)\hat{\mathbf{B}}_\mathbf{i}(\beta)',$$

with $\hat{\mathbf{B}}(\beta) = \{\int_0^\tau [\mathbf{Z}_i(t) - \frac{\mathbf{S}^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}] d\hat{M}_i(\beta,t)\}$ and $d\hat{M}_i(\beta,t) = dN_i(t) - Y_i(t)e^{\beta' \mathbf{Z}_i(t)} d\hat{\Lambda}_0(\beta,t)$. Therefore, this results in a robust sandwich variance estimator $\mathcal{I}(\hat{\beta})^{-1}\hat{\mathbf{\Sigma}}(\beta)\mathcal{I}(\hat{\beta})^{-1}$ for $\hat{\beta}$ (Kalbfleish and Prentice, 2002).

Even though the Andersen-Gill model is the simplest to visualize and set up, it makes the strongest assumption (i.e., independent increment assumption). Since past events are likely to be positively correlated with future events, then when the independence assumption fails it would be best to employ the robust sandwich estimator $n\hat{\mathbf{\Sigma}}(\beta)$ instead of $n^{-1}\mathcal{I}(\beta)$. Cai and Schaubel (2004) pointed out that when the model is approximately true, $\hat{\beta}$ is a useful statistic even when the underlying assumptions do not hold. The Andersen-Gill model has been recommended when the interest is with respect to the overall recurrence rate and when only a small proportion of subjects have $N_i(\tau) \geq 2$ (Lin, 1994).

2.1.1.2. Prentice-Williams-Peterson (PWP) Models

In this subsection the two models proposed by Prentice, Williams and Peterson (1981), which were the first extensions of Cox model for multiple event data, are presented. The intensity function for subject i at time t for the kth recurrence, conditional

on $\mathcal{N}_i(t)$ and on the covariates, can be defined as:

$$\lambda_{ik}(t|\mathcal{N}_i(t), \mathbf{Z}_i(t)) = Y_{ik}(t)\lambda_{0k}(t)e^{\beta_k' \mathbf{Z}_{ik}(t)},$$

and

$$\lambda_{ik}(t|\mathcal{N}_i(t), \mathbf{Z}_i(t)) = Y_{ik}(t)\lambda_{0k}(t - T_{i,k-1})e^{\beta_k' \mathbf{Z}_{ik}(t)},$$

for total and gap times, respectively, with $\mathcal{N}_i = \{N_i(s); s \in [0, t)\}$ denoting the ith subject's event history at time t-. For the risk set indicators, let $Y_{ik}(t) = I(X_{i,k-1} \leq t < X_{i,k})$ and $Y_{ik}(t) = I(X_{i,k} \geq X_{i,k-1}+t)$, respectively, for total time and gap time models. Note that a subject is not at risk for the kth event until he/she has experienced event *k-1*. In both cases, the shape of the baseline hazard function is allowed to be different for different number of preceding events. Hence, this approach produces a stratified proportional intensity model with time-dependent strata, where dependence between event times is accommodated by stratifying on the previous number of occurrences.

The estimation of the regression parameters is carried out through partial likelihood. The estimating equation for the PWP total time model is given by:

$$\mathbf{U}_{TT}(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t) - \frac{\mathbf{Q}_k^{(1)}(\beta_k, t)}{Q_k^{(0)}(\beta_k, t)} \right] dN_{ik}(t),$$

for $k = 1, ..., K,$, where $\mathbf{Q}_k^{(j)}(\beta_k, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}(t)\mathbf{Z}_{ik}(t)^{\otimes j} e^{\beta_k' \mathbf{Z}_{ik}(t)}$ and $N_{ik}(t) = I(T_{i,k} \leq t, \Delta_{ik} = 1)$.

For the PWP gap time model, the estimating equation is given by:

$$\mathbf{U}_{GT}(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t + T_{i,k-1}) - \frac{\mathbf{R}_k^{(1)}(\beta_k, t)}{R_k^{(0)}(\beta_k, t)} \right] d\tilde{N}_{ik}(t),$$

for $k = 1, ..., K$, where $\mathbf{R}_k^{(j)}(\beta_k, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}(t)\mathbf{Z}_{ik}(T_{i,k-1} + t)^{\otimes j} e^{\beta_k' \mathbf{Z}_{ik}(T_{i,k-1}+t)}$ and

$\tilde{N}_{ik}(t) = I(G_{i,k} \leq t, \Delta_{ik} = 1)$.

In both models, it is assumed that the information in $\mathcal{N}_i(t)$ is captured by the covariate vector. In situations where this assumption may be potentially violated, the use of a robust variance estimator is recommended. In their paper, Prentice et al.(1981) suggest that these models are more useful for scenarios of a possibly small number of events on a large number of subjects. When applied to recurrent event data settings, these models generally consider a maximum number of strata, defined such that the risk sets are not so small for latter strata, which could lead to unreliable event-specific estimates. The interpretation of parameters in these models may be limited as the number of events increases. The main problem, as pointed out by some authors (Cai and Schaubel(2004)), is that the assumption of *missing completely at random* (MCAR) is violated because the subjects who have not experienced k events are excluded from the analysis with respect to the *(k+1)*th-event intensity function.

2.1.2 Marginal Hazards Models

2.1.2.1. Wei-Lin-Weissfeld (WLW) Model

In the setting of multivariate time-to-event data, Wei, Lin and Weissfeld(1989) proposed to model the marginal hazard of each failure time using a Cox-type proportional hazards model, such that no specific structure of dependence among the distinct failure times on each subject is imposed. The hazard function for the kth event time of the ith subject assume the form:

$$\lambda_{ik}(t) = \lambda_{0k}(t) e^{\beta_k' \mathbf{Z}_{ik}(t)},$$

for k=1,2,...K. The kth event-specific partial likelihood is given by:

$$PL_k(\beta_k) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta_k' \mathbf{Z}_{ik}(X_{ik})\}}{\sum_{l \in \Re_k(X_{ik})} \exp\{\beta_k' \mathbf{Z}_{lk}(X_{ik})\}} \right]^{\Delta_{ik}},$$

where $\Re_k(t) = \{l : X_{lk} \geq t\}$ is the set of subjects at risk just prior to time t with respect to the kth event time.

The estimator $\hat{\beta}_k$ is defined as the solution to $\mathbf{U}_k(\beta_k) = \mathbf{0}$, where

$$\mathbf{U}_k(\beta_k) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t) - \frac{\mathbf{S}_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} \right] dN_{ik}(t),$$

with $\mathbf{S}_k^{(j)}(\beta_k, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}(t) \mathbf{Z}_{ik}(t)^{\otimes j} e^{\beta_k' \mathbf{Z}_{ik}(t)}$, $Y_{ik}(t) = I(X_{i,k} \geq t)$, $\Delta_{ik} = I(T_{i,k} \leq C_i)$ and $N_{ik}(t) = I(X_{i,k} \leq t, \Delta_{ik} = 1)$.

Under certain regularity conditions, $n^{\frac{1}{2}}(\hat{\beta}_k - \beta_k) \to^{\mathcal{D}} N_p(\mathbf{0}_{p \times 1}, \mathcal{I}_k(\beta_k)^{-1} \mathbf{B}_k(\beta_k) \mathcal{I}_k(\beta_k)^{-1})$, as $n \to \infty$, where a consistent estimator of the asymptotic variance is obtained through

$$\hat{\mathcal{I}}_k(\beta_k) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left[\frac{\mathbf{S}_k^{(2)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} - \left\{ \frac{\mathbf{S}_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} \right\}^{\otimes 2} \right] dN_{ik}(t)$$

and

$$\hat{\mathbf{B}}_k(\beta_k) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_0^{\tau} \left[\mathbf{Z}_{ik}(t) - \frac{\mathbf{S}_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} \right] d\hat{M}_{ik}(\beta_k, t) \right\}^{\otimes 2}$$

with $d\hat{M}_{ik}(\beta_k, t) = dN_{ik}(t) - Y_{ik}(t) e^{\beta_k' \mathbf{Z}_{ik}(t)} d\hat{\Lambda}_{0k}(\beta_k, t)$. The robust sandwich covariance matrix defined above accounts for the dependence of the multiple failure times.

The cumulative baseline hazard function for the kth event is estimated by $d\hat{\Lambda}_{0k}(\beta_k, t) = n^{-1} \int_0^t dN_{\cdot k}(s) / S_k^{(0)}(\hat{\beta}_k, s)$, where $dN_{\cdot k}(t) = \sum_{i=1}^{n} dN_{ik}(t)$.

The inferences regarding β_k are valid asymptotically, irrespective of the true intra-subject correlation structure. However there is some debate in the literature regarding the appropriateness of WLW model for recurrent data, especially due to the interpre-

tation of regression coefficients. Two main issues have been discussed in applying this approach to recurrent event settings: (i) the possibility of a subject to be at risk for the *(k+1)*th event prior to having experienced the kth event (Cook and Lawless, 1997); (ii) a carry-over effect, which leads to an overestimation of regression coefficients (Kelly and Lim, 2000). Hence, WLW is mostly recommended to situations where there are different types of events for the same subject instead of recurrent events.

2.1.2.2. Lee-Wei-Amato (LWA) Model

Another marginal approach is the LWA model (Lee, Wei and Amato, 1992), which considers highly stratified data. The marginal hazard for the kth event time of the ith subject has the form:

$$\lambda_{ik}(t) = \lambda_0(t) e^{\beta' \mathbf{Z}_{ik}(t)},$$

where an unspecified common baseline hazard function $\lambda_0(t)$ is assumed. Let $\beta = (\beta_1, ...\beta_p)'$ be the common regression parameter among K marginal models . Under the assumption that $(\mathbf{X}_i, \mathbf{\Delta}_i, \mathbf{Z}_i(t))$, $i = 1, .., n$ are independent and identically distributed with bounded $\mathbf{Z}_{ik}$ (Lee, Wei & Amato, 1992), the parameter estimates are obtained through the maximization of the pseudo-partial likelihood:

$$PL(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\frac{\exp\{\beta' \mathbf{Z}_{ik}(X_{ik})\}}{\sum_{l=1}^{n} \sum_{m=1}^{K} Y_{lm}(X_{ik}) \exp\{\beta' \mathbf{Z}_{lm}(X_{ik})\}} \right]^{\Delta_{ik}},$$

where $Y_{ik}(t) = I(X_{ik} \geq t)$.

The corresponding score equations are given by:

$$\mathbf{U}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^{\tau} \left[\mathbf{Z}_{ik}(t) - \frac{\mathbf{S}^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right] dN_{ik}(t),$$

where $\mathbf{S}^{(j)}(\beta, t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} Y_{ik}(t) \mathbf{Z}_{ik}(t)^{\otimes j} e^{\beta' \mathbf{Z}_{ik}(t)}$.

Lee, Wei and Amato (1992) showed that, under certain regularity conditions, the resulting estimators $n^{\frac{1}{2}}(\hat{\beta}_1' - \beta_1', ..., \hat{\beta}_p' - \beta_p')'$ are asymptotically jointly normal with mean zero and covariance matrix that can be consistently estimated by the sandwich type covariance estimator.

This approach assumes that the number of subjects is much larger than the number of events. A major concern about the use of such approach for recurrent data is that the LWA model allows a subject to be at risk for several events simultaneously. That is, a subject with j risk intervals may contribute to a risk set j times (Kelly and Lim, 2000). Simulation studies also suggest the same carry over effect as observed with WLW model.

2.1.3 Frailty Models

In the last several years there has been significant research concerning the addition of random effects to survival models (Clayton and Cuzick, 1985; Hougaard, 1986; McGilchrist and Aisbett, 1991; Klein, 1992; Duchateau et al, 2003). In this setting, the hazard function for each individual may depend on observed risk variables but usually not all such variables are known or measurable. This unknown factor is usually called individual heterogeneity or frailty. For the recurrent event data setting, the dependence of repeated measures (i.e., the recurrence times) is modelled through the introduction of a common random effect (i.e., frailty) for each individual. Considering a Cox proportional hazard model for recurrent event data including a multiplicative heterogeneity or frailty term W_i, i=1,...,n, the hazard function has the form:

$$\lambda_{ik}(t|W_i) = w_i \lambda_0(t) e^{\beta' \mathbf{z}_{ik}(t)},$$

for $i = 1, ..., n$ and $k = 1, ...K_i$. The frailty terms are assumed to be independent and

with a common frailty density. The most commonly used frailty distribution is the gamma distribution with mean 1 and variance θ, such that:

$$f_{W_i}(w) = \frac{w^{1/\theta - 1} \exp(-w/\theta)}{\theta^{1/\theta}\Gamma(\frac{1}{\theta})}, \theta > 0, w > 0$$

Even though the assumption of gamma distribution for the frailty terms has been commonly used, this may not be always plausible. Thus, other distributions for the frailty terms has been also proposed, such as positive stable distribution (Hougaard, 1986) and lognormal distribution (Hougaard, 2000).

Let t_{ik1} be the start of the k^{th} at-risk period and t_{ik2} the end of the k^{th} at-risk period. The partial likelihood function (without ties) for the model above is given by:

$$\prod_{i=1}^{n}\prod_{k=1}^{K_i}\left[\frac{W_i \exp(\beta' \mathbf{Z}_{ik})}{\sum_{l=1}^{n} Y_l(t_{ik2}) W_l \exp(\beta' \mathbf{Z}_{lk})}\right]^{\Delta_{ik}},$$

where $Y_l(t_{ik2}) = I(X_{ik} \geq t_{ik2})$.

The parameter estimates for this model are obtained through the EM algorithm, making use of the partial likelihood expression in the maximization step as showed by Klein (1992). An alternative approach is to use a penalized partial likelihood for the estimation of the shared frailty, which consider the general framework for penalized survival models along with its application to smoothing splines (Therneau and Grambsch, 2001). McGilchrist and Aisbett (1991), on the other hand, discussed the use of a Newton-Raphson procedure to estimate β and the vector of frailties $\mathbf{w}$. However, this approach may not converge if the major variation of the risk variables is from subject to subject rather than within each subject.

Duchateau et al (2003) also consider parametric frailty models for recurrent event data. Their models incorporate inter-subject heterogeneity through a random effect and specify the functional form of the baseline hazard. One of the parametric models

discussed in that paper defines a distinct Weibull baseline hazard for the first event and for the subsequent events, considering that the first event is different in nature from the subsequent events, such that:

$$\lambda_i(t) = \lambda_f \gamma_f t^{\gamma_f - 1} W_i \exp(\beta' \mathbf{Z}_i),\ 0 \leq t \leq t_{i12}, \text{ for the first event}$$

and

$$\lambda_i(t) = \lambda_g \gamma_g (t - t_{ik1})^{\gamma_g - 1} W_i \exp(\beta' \mathbf{Z}_i),\ t_{ik1} \leq t \leq t_{ik2},\ k = 2, .., K_i$$

The parameter estimates for this model can be obtained by maximizing the observable likelihood based on this hazard function.

Some debate in the literature regarding the use of frailty models are related to (i) the possibility of mis-specification of the dependence structure and (ii) the amount of information, such as total number of events, number of groups, and the distribution of events/group, required to produce stable frailty estimates. According to Hougaard (2000), there is no single family of frailty distribution having all desirable properties. Therefore, the choice of the frailty distribution requires more detailed study of the model properties of each distribution family and of which properties are relevant to the actual problem considered.

2.1.4 Marginal Means and Rates Models

One interesting approach for recurrent events in survival analysis is modelling marginal means and rates. Pepe and Cai (1993) proposed modelling the rate functions $r_{ik}(t)$, which represent the average intensity or rate among those who have experienced *(k-1)* events. Using such approach, they avoided the problem of being at risk of having the kth event without having experienced the *(k-1)*th event. Later, Lawless and Nadeau (1995) studied the estimation of regression coefficients for marginal means/rates models,

primarily considering the discrete time case. More recently, Lin et al (2000) provided a rigorous justification for the marginal means/rates models and develop inference procedures for the continuous time setting, which are presented here.

As Cai and Schaubel (2004) pointed out, one of the main appeals of using such approach is that the mean number of events is usually of direct interest of investigators and is also easier to be understood, especially for non-statisticians. Using counting process notation, the rate function can be always defined as $d\mu_i(s) = E[dN_i(s)|\mathbf{Z}_i(s)]$. On the other hand, $\mu_i(t) = \int_0^t d\mu_i(s)$ can only be interpreted as a mean function when $\mathbf{Z}_i(.)$ consists of time invariant or external time-dependent covariates, in which case they are unaffected by the recurrent event process. If $\mathbf{Z}_i(.)$ includes time-dependent internal covariates, then $\mu_i(t)$ can only be interpreted as a cumulative rate function.

The proportional rates models considered by Lin et al(2000) may be defined as:

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

Thus, $\mu_i(t) = \int_0^t \exp\{\beta' \mathbf{Z}_i(s)\} d\mu_0(s)$. In the case that $\mathbf{Z}_i$ represent time invariant covariates, then $\mu_i(t) = \mu_0(t) \exp\{\beta' \mathbf{Z}_i\}$ is referred to as the proportional means models, with $\mu_0(t)$ being an arbitrary baseline mean function and β unknown regression parameters. According to Lin et al (2000), the proportional rates model characterizes the rate of the counting process under the Andersen-Gill intensity model. The A-G model implies a proportional rates model with $d\mu_0(t) = \lambda_0(t)dt$, but the converse is not true. The proportional rates model is more versatile than the A-G model in the way that it allows arbitrary dependence structure among recurrent events.

For the developments provided by Lin et al(2000), some regularity conditions should be considered, such that they assume a finite follow-up interval $[0, \tau]$ with $P(C_i \geq \tau) > 0$. They assume that $\{N_i(.), Y_i(.), \mathbf{Z}_i(.)\}$, $i = 1, ..., n$ are independent and identically

distributed, where $Y_i(t) = I(C_i \geq t)$. They assume that $N_i(\tau)$, $i = 1, ..., n$, are bounded by a constant, that $\mathbf{Z}_i(.)$, $i = 1, ..., n$, have bounded total variation and finally that the information matrix $\mathcal{I}_n(\beta) = -\partial \mathbf{U}(\beta)/\partial \beta'$ converges in probability to a positive definite matrix as $n \to \infty$.

Thus, the inference on the regression parameters is defined by the solution of the partial likelihood score function for β, e.g., $\mathbf{U}_n(\beta, \tau) = \mathbf{0}_{p\times 1}$, such that:

$$\mathbf{U}_n(\beta, t) = \sum_{i=1}^{n} \int_0^t \left[\mathbf{Z}_i(\mathrm{u}) - \frac{\mathbf{S}^{(1)}(\beta, u)}{S^{(0)}(\beta, u)} \right] dN_i(u),$$

where $S^{(0)}(\beta, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{\beta' \mathbf{Z}_i(t)\}$ and
$\mathbf{S}^{(1)}(\beta, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{\beta' \mathbf{Z}_i(t)\} \mathbf{Z}_i(t)$. The baseline mean $\mu_0(t)$ is estimated by the Breslow-type estimator, such that $\hat{\mu}_0(t) = n^{-1} \int_0^t dN_.(u)/S^{(0)}(\beta, u)$.

Under the regularity conditions presented above, Lin et al (2000) showed that, under the proportional rates model, $n^{-\frac{1}{2}} \mathbf{U}(\beta, t)$ asymptotically follows a multivariate normal distribution with zero mean. The covariance function between time points s and t is given by $\mathbf{B}(\beta, s, t)$, where
$\mathbf{B}(\beta, s, t) = E[\int_0^s \{\mathbf{Z}_i(r) - \frac{\mathbf{S}^{(1)}(\beta,r)}{S^{(0)}(\beta,r)}\} dM_i(r) \times \int_0^t \{\mathbf{Z}_i(r) - \frac{\mathbf{S}^{(1)}(\beta,r)}{S^{(0)}(\beta,r)}\} dM_i(r)]$, for $0 \leq s$, $t \leq \tau$ with $dM_i(t) = dN_i(t) - Y_i(t) \exp\{\beta' \mathbf{Z}_i(t)\} d\mu_0(t)$. Lin et al.(2000) also showed that, under the proportional rates model, $n^{\frac{1}{2}}(\hat{\beta} - \beta) \to^{\mathcal{D}} N(\mathbf{0}_{p\times 1}, \mathcal{I}(\beta)^{-1} \mathbf{B}(\beta) \mathcal{I}(\beta)^{-1})$, where $\mathcal{I}(\beta)$ is the limiting value of $\mathcal{I}_n$ evaluated at β and
$\mathbf{B}(\beta)$=$E[(\int_0^\tau \{\mathbf{Z}_i(\mathrm{s}) - \mathbf{S}^{(1)}(\beta, s)/S^{(0)}(\beta, s)\} dM_i(s))^{\otimes 2}]$.

In order to test $H_o : \beta = \mathbf{0}$, nonparametric statistics $\mathbf{U}'(\mathbf{0}) \mathbf{B}^{-1}(\mathbf{0}) \mathbf{U}(\mathbf{0})$ can be used.

2.1.5 Nonparametric Estimation

2.1.5.1. Nonparametric Estimation for Recurrent Event Data

The problem of nonparametric estimation for recurrent event data has, among oth-

ers, been considered by Pepe and Cai (1993), Lawless and Nadeau (1995), Cook and Lawless (1997) and Wang and Chiang (2002). All methods studied by these authors are formulated to time-to-event data and model the occurrence rate of recurrent events in a specified time interval $[0, \tau_i]$.

Pepe and Cai (1993) proposed to display rate functions that distinguish first (k=1) and recurrent events (k=2,...), such that they are, respectively, defined as

$$r_{i1}(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P\left[\text{event in}(t, t+\Delta) | \text{ at risk and no previous event at t}\right]$$

and

$$r_{ik}(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P\left[\text{event in}(t, t+\Delta) | \text{ at risk and (k-1) previous events at t}\right] \text{ for k=2,...}$$

Similar to the marginal approach, Pepe and Cai (1993) proposed to leave the intra-subject correlation unspecified. Note, however, that the rate functions r_{ij}, k=1,2..., are conditional on having experienced (k-1) events.

The methods discussed by Lawless and Nadeau (1995) and Cook and Lawless (1997), on the other hand, focus on a discrete time framework. Consider recurrent event processes $\{N_i(t) : t \geq 0\}$, i=1,...,k, that are independent and have the same cumulative mean function, which can be defined as $CM(t) = E\{N_i(t)\}$. Let $n_i(t) \geq 0$ represent the number of events that occur at time t for the ith subject. Hence, the mean function can be defined as $m(t) = E\{n_i(t)\}$ and, consequently, $CM(t) = \sum_{s=0}^{t} m(s)$. Define $\delta_i(t) = 1$ if $t \leq \tau_i$ and 0 otherwise. If the $n_i(t)'s$ are independent Poisson random variables with means m(t) then the maximum likelihood estimate (MLE) of m(t) is given by $\hat{m}(t) = n_{\cdot}(t)/\delta_{\cdot}(t)$, where $n_{\cdot}(t) = \sum_{i=1}^{n} \delta_i(t) n_i(t)$ and $\delta_{\cdot}(t) = \sum_{i=1}^{n} \delta_i(t)$ denote, respectively, the total number of events and the total number of subjects at risk at time t.

Therefore, the estimate of CM(t), for $0 \leq t \leq \tau = max(\tau_i)$, is

$$\hat{CM}(t) = \sum_{s=0}^{t} \frac{n_{\cdot}(s)}{\delta_{\cdot}(s)} = \sum_{i=1}^{n} \sum_{s=0}^{t} \frac{\delta_i(s) n_i(s)}{\delta_{\cdot}(s)},$$

which is known as the Nelson-Aalen estimator and it is well known as a nonparametric MLE for the cumulative intensity in counting process models (Lawless and Nadeau, 1995). It was also shown by these authors that the variance of $\hat{CM}(t)$ is estimated consistently in large samples by:

$$\hat{V}(t) = \sum_{i=1}^{n} \left\{ \sum_{s=0}^{t} \frac{\delta_i(s)}{\delta_{\cdot}(s)} [n_i(s) - \hat{m}(s)] \right\}^2,$$

In the continuous case, the nonparametric estimators may be written as $\hat{CM}(t) = \int_0^t dN_{\cdot}(s)/\delta_{\cdot}(s)$ and $\hat{V}(t) = \sum_{i=1}^{n} \{\int_0^t \delta_i(s)/\delta_{\cdot}(s)[dN_i(s) - dN_{\cdot}(s)/\delta_{\cdot}(s)]\}^2$, where $dN_i(s)$ represents the number of events for subject i at time s and $dN_{\cdot}(s) = \sum_{i=1}^{n} dN_i(s)$. The estimates $\hat{CM}(t)$ and $\hat{V}(t)$ are robust because they are simple moment estimates (Lawless and Nadeau, 1995).

In order to test whether the cumulative mean function among two groups of subjects are equal, an approach proposed by Lawless and Nadeau (1995) may be applied, using the statistic defined as

$$U = \sum_{s=0}^{\tau} \frac{\delta_{0\cdot}(s)\delta_{1\cdot}(s)}{\delta_{0\cdot}(s) + \delta_{1\cdot}(s)} \left[\frac{n_{1\cdot}(s)}{\delta_{1\cdot}(s)} - \frac{n_{0\cdot}(s)}{\delta_{0\cdot}(s)} \right],$$

where $n_{j\cdot}(s)$ is the number of events in group j at time s and $\delta_{j\cdot}(s)$ is the number of subjects under observation in group j at time s (j=0,1). It follows that a robust variance estimate for U is given by

$\hat{var}(U) = \sum_{j=0}^{1} \sum_{i=1}^{n_j} \{\sum_{s=0}^{\tau} \delta_{ji}(s) \times \frac{[\delta_{\cdot\cdot}(s) - \delta_{j\cdot}(s)]}{\delta_{\cdot\cdot}(s)} [n_{ji}(s) - \frac{n_{j\cdot}(s)}{\delta_{j\cdot}(s)}]\}^2$, where $n_{ji}(s)$ is the number of events at time s for the ith subject of group j (j=0,1), $\delta_{ji}(s)$ indicates whether

the ith subject of group j is observed at time s and the dots indicate summation over the appropriate indices.

To test $H_0 : CM^{(0)}(t) = CM^{(1)}(t)$, for $0 \leq t \leq \tau$, the test statistic $Z = U/\hat{var}(U)^{1/2}$ may be used. However, Lawless and Nadeau (1995) pointed out that this statistics is only effective at detecting differences when the cumulative mean function of the two groups are proportional or roughly so. If other types of departures from equality are expected, they proposed the use of general weight functions w(s) on the test. Cook and Lawless (1997) discussed similar robust tests for no treatment effect.

Wang and Chiang (2002), on the other hand, focus on the discussion of nonparametric procedures for the estimation of the cumulative occurrence rate function (CORF) and the occurrence rate function (ORF). The estimator for the CORF is the same as that already discussed by Lawless and Nadeau (1995). However, they used smoothing techniques for the estimation of the ORF. Wang and Chiang (2002) assume that, for the ith subject, $N_i(t)$ is distributed as a non-stationary Poisson process with the subject-specific intensity function $\lambda_i(t)$, where $N_i(t)$ denote the number of recurrent events occurring at or prior to t, $t \geq 0$. Let the ORF of recurrent events for the target population to be defined as $\lambda(t) = E[\lambda_i(t)]$. Thus, the kernel estimator of the subject-specific intensity function $\lambda_i(t)$ for t in the interval $[0, C_i]$, where $C_i(t)$ is the censoring time at which the observation of the recurrent events is terminated, is defined by

$$\hat{\lambda}_i(t) = \sum_{l=1}^{e_i} K_{c_i}\left(\frac{t - t_{il}}{h}\right),$$

where e_i defines the index for the last event occurring at or prior to C_i, $K_{c_i}(.)$ is a boundary kernel density of Gasser and Müller (1978) with adjustment for the censoring time C_i, and h is a positive-valued bandwidth. Therefore, the kernel estimator for $\lambda(t)$, say $\hat{\lambda}_h(t)$, is given by averaging the subject-specific estimators of subjects who are still

at risk at t, such that

$$\hat{\lambda}_h(t) = \sum_{i=1}^{n} \left(\frac{I(C_i \geq t)}{n_R(t)} \right) \hat{\lambda}_i(t),$$

$t \in [0, \tau]$, with $n_R(t)$ being the number of subjects in the risk set $R(t) = \{i : C_i \geq t\}$.

2.1.5.2. Nonparametric Estimation of Gap Time Distributions

Another way of describing and modelling the recurrent events is through gaps or waiting times between successive events. Such approaches are often useful when predictions concerning time to the next event are of interest. However, a major difficulty when dealing with gap time distributions of several events is the induced dependent censoring, which results from the lack of within-subject gap time independence (Lin, Sun and Ying, 1999; Lin and Ying, 2001; Schaubel and Cai, 2004).

Lin et al (1999) propose a nonparametric approach for estimating the joint distribution of the gap times without imposing any assumption on the dependence structure of the gap times. As usual, assume that C is independent of the failure (total) times $T_1, ..., T_k$. However, for any k=2,...,K, the gap time G_k is subject to right censoring by $C - T_{k-1}$, which is naturally correlated with G_k unless G_k is independent of T_{k-1}.

Considering K=2 and that there are n independent subjects in the study, let F_1 and F_2 be the marginal distribution functions of G_1 and G_2 and let F be their joint distribution function, such as:

$$F_1(t) = Pr(G_1 \leq t), F_2(t) = Pr(G_2 \leq t)$$

and

$$F(t_1, t_2) = Pr(G_1 \leq t_1, G_2 \leq t_2) = H(t_1, 0) - H(t_1, t_2),$$

where $H(t_1, t_2) = Pr(G_1 \leq t_1, G_2 > t_2)$.

Let $G_{i1}^* = G_{i1} \wedge C_i$ and $G_{i2}^* = (G_{i1} + G_{i2}) \wedge C_i$. The authors suggest the estimator

$$\hat{H}(t_1, t_2) = n^{-1} \sum_{i=1}^{n} \frac{I(G_{i1}^* \leq t_1, G_{i2}^* - G_{i1}^* > t_2)}{\hat{\tilde{S}}(G_{i1}^* + t_2)},$$

where $\hat{\tilde{S}}(t)$ is the Kaplan-Meier estimator of the survival function of the censoring time variable, i.e. $\tilde{S}(t) = P(C > T)$, based on the data $(G_{i1}^*, 1 - \Delta_{i1})$ or $(G_{i2}^*, 1 - \Delta_{i2})$, $i = 1, ...n$. The estimator $\hat{H}(t_1, t_2)$, as well as any other estimator, is confined to $\{(t_1, t_2) : t_1 + t_2 < \tau_c\}$, where $\tau_c = sup\{P(C_i \geq t) > 0\}$, as a result of the estimability constraint.

Hence, the joint distribution function is estimated by $\hat{F}(t_1, t_2) = \hat{H}(t_1, 0) - \hat{H}(t_1, t_2)$ and it is possible to estimate the conditional distribution function $F_{2|1}(t_2|t_1) = P(G_2 \leq t_2|G_1 \leq t_1)$ through $\hat{F}_{2|1}(t_2|t_1) = 1 - \hat{H}(t_1, t_2)/\hat{H}(t_1, 0)$.

The authors showed that the estimator $\hat{F}(t_1, t_2)$ is strongly consistent, and the process $n^{\frac{1}{2}}\{\hat{F}(t_1, t_2) - F(t_1, t_2)\}$ converges weakly to a bivariate zero-mean Gaussian process with a covariance function that can be estimated using empirical quantities, while $\hat{F}_{2|1}(t_2|t_1)$ is also strongly consistent and $n^{\frac{1}{2}}\{\hat{F}_{2|1}(t_1|t_2) - F_{2|1}(t_1|t_2)\}$ converges weakly to zero-mean Gaussian process with the covariance function as showed in Lin et al (1999).

Since it is often of interest to compare the gap distributions between two or more groups, Lin and Ying (2001) propose nonparametric tests that allow such comparison considering the estimators defined above. These authors also suggest a class of statistics to test the independence between two or more gap times.

A more recent nonparametric conditional estimator was proposed by Schaubel and Cai (2004) as an alternative to existing methods. The authors suggest to estimate the conditional survival function $S_k(t, t_{k-1}) \equiv P(G_{ik} > t|G_{i,k-1} \leq t_{k-1})$, for $t \in [0, \tau_k]$ with $t_{k-1} + \tau_k \leq \tau_c$, through its relationship with the corresponding cumulative hazard

function $\Lambda_k(t; t_{k-1})$, where $S_k(t, t_{k-1}) = e^{-\Lambda_k(t;t_{k-1})}$.

Let $\tilde{G}_{i,k} = G_{i,k} \wedge \tilde{C}_{i,k}$ denote the observed gap times and $\tilde{C}_{i,k} = C_i - X_{i,k-1}$ denote the gap censoring times. Thus, the proposed estimator for the cumulative hazard function has the form:

$$\hat{\Lambda}_k(t; t_{k-1}) = n^{-1} \sum_{i=1}^{n} \int_0^t \frac{\hat{W}_{ik}(s)}{R_k(s)} N_{ik}(ds; t_{k-1}),$$

where $\hat{W}_{ik}(s) = Y_{ik}(s; t_{k-1}) \tilde{S}(s + T_{i,k-1})^{-I(k \geq 2)}$; with $Y_{ik}(s; t_{k-1}) = I(\tilde{G}_{ik} \geq s, T_{i,k-1} \leq t_{k-1})$; $\hat{\tilde{S}}(t)$ being the Kaplan-Meier estimator of $\tilde{S}(t) = P(C > T)$ based on $(X_{i,K}, 1 - \Delta_{iK})$, $i = 1, ..., n$; $R_k(s) = n^{-1} \sum_{i=1}^{n} \hat{W}_{ik}(s)$ and $N_{ik}(s; t_{k-1}) = I(\tilde{G}_{ik} \leq s, \Delta_{ik} = 1, T_{i,k-1} \leq t_{k-1})$.

Hence, the conditional survival function can be estimated by $\hat{S}_k(t, t_{k-1}) = e^{-\hat{\Lambda}_k(t;t_{k-1})}$ for $t \in [0, \tau_k]$. Furthermore, the authors also show that $n^{\frac{1}{2}}\{\hat{\Lambda}(t_k, t_{k-1}) - \Lambda(t_k, t_{k-1})\}$ converges weakly to a zero-mean Gaussian process with a covariance function that can be estimated using empirical processes.

2.2 Methods considering Dependent Censoring

In the previous sections we reviewed approaches that assume independent censoring, e.g, the censoring process is unrelated to the event failure process. However, according to Ghosh and Lin (2003) the recurrent event times in a typical medical study are often subject to both independent and dependent censoring. The dependent or informative censoring arises when the censoring time depends on the observed or unobserved recurrent event times. This would happen if, for instance, the subjects who are at higher risk of recurrent events tend to be withdrawn from the study earlier. In such scenario the subjects can potentially experience further events after the censoring time, but these events will not be observed by the investigators. Another form of dependent

censoring would occur because of terminal events, such as death. In the case of informative censoring the *ad hoc* estimation procedures from the observed data will result in inconsistent estimators (Miloslavsky et al, 2004). Thus in order to take into account the potential dependent censoring in the analysis of recurrent event data we review in this section two approaches that were proposed recently to handle dependent censoring (Wang, Qin and Chiang, 2001; Miloslavsky et al, 2004).

2.2.1 Wang, Qin and Chiang (WQC) Model

Wang, Qin and Chiang (2001) proposed to model the occurrence of recurrent events by a subject-specific nonstationary Poisson process via a latent variable, allowing the censoring mechanism be possibly informative. The distribution of both the censoring and latent variables are treated as nuisance parameters.

Let $N(t)$ be the number of recurrent events at or before t, $t \geq 0$, and suppose that the occurrence rate of recurrent events in the interval $[0, \tau]$ is of interest. Thus, in order to explore the association between the covariates $\mathbf{Z}$ and N(.), consider a multiplicative intensity model and assume the following conditions:

(i) There exists a nonnegative valued latent variable U so that, conditioning on $(\mathbf{Z}, u)$, $N(t)$ is a nonstationary Poisson process with the intensity function $u\lambda_0(t)\exp\{\beta'\mathbf{Z}\}$, where β is a $p \times 1$ vector of parameters and the baseline intensity $\lambda_0(t)$ is a continuous function. The latent variable satisfies $E[U|\mathbf{Z}] = 1$. This assumption implies the marginal proportional rate function $\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\beta'\mathbf{Z}\}$, which is the focus of interest for the regression inferences.

(ii) Conditioning on $(\mathbf{Z}, u)$, N(.) is independent of C, which denotes the censoring time.

Define the density function

$$f(t) = \frac{\lambda_0(t) I(0 \leq t \leq \tau)}{\Lambda_0(\tau)} = \frac{u_i \lambda_0(t) I(0 \leq t \leq \tau)}{u_i \Lambda_0(\tau)},$$

as the normalized function for both the baseline intensity $\lambda_0(t)$ and the subject-specific intensity, $u_i \lambda_0(t)$, when $u_i > 0$, such that $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ denotes the baseline cumulative rate function. This density function, or the corresponding cumulative distribution function F(t), can be thought of as shape function for the model. When considering the estimation of the parameter β of the marginal proportional rate function, the density function may be expressed as

$$f(t) = \frac{\lambda_0(t) I(0 \leq t \leq \tau)}{\Lambda_0(t)} = \frac{u_i \lambda_0(t) \exp\{\beta' \mathbf{Z}\} I(0 \leq t \leq \tau)}{u_i \Lambda_0(t) \exp\{\beta' \mathbf{Z}\}},$$

which remains as the shape function for the subject-specific intensity, $u_i \lambda_0(t) \exp\{\beta' \mathbf{Z}\}$, when $u_i > 0$.

Let k_i denote the index for the last event occurring at or prior to C_i. A conditional likelihood will be used in the estimation procedures instead of the full likelihood because of the involvement of the nuisance parameters. For subject i, conditional on (C_i, u_i, k_i), the event times $(T_{i1}, T_{i2}, ..., T_{ik_i})$ are the order statistics of a set of iid random variables with the density function $f(t) I(0 \leq t \leq C_i) / F(C_i)$ (Ross, 1983). Thus, the conditional likelihood function can be derived as

$$\begin{aligned} L_c &= \prod_{i=1}^{n} p(T_{i1}, T_{i2}, ..., T_{ik_i} | C_i, u_i, k_i) \\ &= \prod_{i=1}^{n} \left\{ k_i! \prod_{j=1}^{k_i} \frac{f(T_{ij})}{F(C_i)} \right\} \propto \prod_{i=1}^{n} \prod_{j=1}^{k_i} \frac{f(T_{ij})}{F(C_i)}. \end{aligned}$$

The conditional likelihood L_c involves then only the shape function F (or f) and does not require information on the unobserved $\{u_i\}$. The likelihood function L_c is a particular case of nonparametric likelihood for right-truncated data where the truncation time for T_{ij} is C_i. Under the model, the distribution function F(t) is unknown, and thus can be estimated by the nonparametric maximum likelihood estimator $\hat{F}$(t). Under regularity conditions, the estimator $\hat{F}(t)$ is known to have a simple product-limit representation, such that $\hat{F}(t) = \prod_{s_{(l)}>t}(1 - d_{(l)}/N_{(l)})$, where $\{s_{(l)}\}$ are the ordered and distinct values of the event times $\{T_{ij}\}$, $\{d_{(l)}\}$ is the number of events occurring at $s_{(l)}$, and $\{N_{(l)}\}$ is the total number of events with event time and censoring time satisfying $T_{ij} \leq s_{(l)} \leq C_i$.

A class of estimating equations for $\gamma' = (\ln \beta_0, \beta')$ is defined as

$$n^{-1} \sum_{i=1}^{n} w_i \bar{\mathbf{Z}}_i' (k_i \hat{F}^{-1}(C_i) - e^{\gamma' \bar{\mathbf{z}}_i}) = 0,$$

where $\bar{\mathbf{Z}}_i = (1, \mathbf{Z}_i)$, $\beta_0 = \Lambda_0(\tau)$ and w_i is a weight function depending on $(\mathbf{Z}_i, \gamma, F)$. Wang, Qin and Chiang (2001) showed that the solution of this class of estimating equations has the property that $\sqrt{n}(\hat{\gamma} - \gamma)$ converges weakly to a multivariate normal distribution with zero mean and covariance matrix which can be consistently estimated if the marginal rate model is correctly specified.

2.2.2 IPCW Models

Inverse probability of censoring weighted(IPCW) estimators for the regression parameters in the Andersen-Gill model and in the proportional rates model were proposed by Miloslavsky et al (2004) in order to obtain consistent estimators in the full data models from the observed data in the presence of dependent censoring. Let $\mathbf{V} = \bar{V}(\tau)$, where $\bar{V}(\tau) = \{V(s) : s \leq \tau\}$, stands for everything that can be observed on a randomly

selected subject in the interval $(0, \tau]$ if the subject is not subject to censoring. In particular, define $\bar{V}(\tau) = \{N(\tau), \mathbf{Z}(\tau)\}$, where $\bar{N}(\tau) = \{N(s) : s \leq t\}$, $N(t) = \sum_k I(T_k \leq t)$ is the recurrent event counting process of interest and τ refer to the end time of the study. Consider the following multiplicative intensity model of interest:

$$E\{dN(t)|\bar{\mathbf{W}}(t-)\} = Y_\lambda(t)\lambda(t) = Y(t)\lambda_0(t)\exp\{\beta'\gamma(t)\}$$

where $\bar{\mathbf{W}}(t) = \{\bar{N}(t), \mathbf{Z}^*(t)\}$, with $\mathbf{Z}^*(t) \subset \mathbf{Z}(t)$, consisting of part of the full data covariate process $\mathbf{Z}(t)$ and $\gamma(t)$ is a function of $\bar{\mathbf{W}}(t-)$.

Often the full data is not observed but their censored version. Denote the observed data random variable by $\mathbf{E}$=$\{min(\tau, C), \Delta^* = I(\tau < C), \bar{V}(\tau \wedge C)\}$ and let $A(t) = I(C < t)$ be the censoring process, where $C = \infty$ if C is censored by τ. The distribution of the observed data $\mathbf{E}$ is indexed by the full data distribution F_V and by the conditional distribution $G(.|V)$ of the censoring variable C given V. Refer to $G(.|V)$ as the censoring mechanism. Thus, the conditional hazard of the censoring mechanism given the full data $\mathbf{V}$ is $\lambda_c(t|V) = E\{dA(t)|\bar{A}(t-) = 0, V\}$.

The full data parameter of interest is not identifiable from the distribution of the observed data if the censoring mechanism is allowed to depend on unobserved components of $\mathbf{V}$. Hence *coarsening at random* (CAR) is assumed, which implies that given the full data $\mathbf{V}$=v, the censoring event defining the observed data $\mathbf{E}$=e depends only on the observed part of v. This methodology requires a model for the censoring mechanism, which can be given by:

$$\lambda_c(t|\bar{V}(t-)) = Y_c(t)\lambda_{0,c}(t)\exp\{\beta_c\xi_c(t)\},$$

where $Y_c(t)$ is the at-risk indicator for censoring, $\lambda_{0,c}(t)$ is an unspecified baseline hazard

and $\xi_c(t)$ is a known function of $\bar{V}(t-)$. Under CAR the intensity of observed data process reduces to the intensity of the full data counting process if the conditioning set of the full data intensity model includes the whole past $\bar{V}(t)$.

The class of all full data estimating function for the multiplicative intensity model of interest is given by $\{D_h(.|\beta,\lambda_0) = \int[h\{t,\bar{\mathbf{W}}(t-)\} - g(h)(t)]dM_{\beta,\lambda_0(t)} : h\}$, where $g(h)(t) = g(t) = \frac{E[h\{t,\bar{\mathbf{W}}(t-)\}Y(t)\exp\{\beta\gamma(t)\}]}{E[Y(t)\exp\{\beta\gamma(t)\}]}$, $dM_{\beta,\lambda_0}(t) = dN(t) - E\{dN(t)|\bar{\mathbf{W}}(\mathbf{t}-)\}$ and h is a user-defined function (van der Laan and Robins, 2002).

Since the marginal A-G multiplicative intensity model of interest is not conditioning on $V(t)$, but only on some subset $\bar{\mathbf{W}}(t)$, the authors proposed estimating equations for the parameter of interest in this general model by using IPCW mapping (Robbins and Rotnitzky, 1992), for which the main idea is to map full data estimating functions into observed data estimating functions.

Let $\tilde{\Delta}^*(t) = I(C > t)$ and $h^*\{t,\bar{\mathbf{W}}(t-)\} = h\{t,\bar{\mathbf{W}}(t-)\} - g(h)(t)$. Then the choice of IPCW estimating function is given by:

$$U_G(\mathbf{E}|D_h) = \int_0^\tau h^*\left\{t,\bar{\mathbf{W}}(t-)\right\}\frac{dM_{\beta,\lambda_0}(t)\tilde{\Delta}^*(t)}{\bar{G}(t|\mathbf{V})},$$

where $\bar{G}(t|\mathbf{V}) = P(C > t|\mathbf{V})$. Note that $U_G(.|D_h)$ satisfies $E\{U_G(\mathbf{E}|D_h)|\mathbf{V}\} = D_h(\mathbf{V}|\beta,\lambda_0)$ under the assumption that $P(C > \tau|\mathbf{V}) > \delta > 0$, for some $\delta > 0$ and hence it yields consistent estimators in the presence of dependent censoring.

A particular choice for $h^*\{t|\bar{\mathbf{W}}(t-)\}$ is

$$h^*\{t|\bar{\mathbf{W}}(t-)\} = \left(\gamma(t) - \frac{E[\gamma(t)\bar{G}\{t|\bar{\mathbf{W}}(t-)\}Y(t)\exp\{\beta\gamma(t)\}]}{E[\bar{G}\{t|\bar{\mathbf{W}}(t-)\}Y(t)\exp\{\beta\gamma(t)\}]}\right)\bar{G}\{t|\bar{\mathbf{W}}(t-)\}$$

Applying the time-dependent weighting to the full data estimating equation yields

the following observed data estimating equation:

$$\mathbf{U}_G(\mathbf{E}|D_h^*) = \int_0^\tau \left(\gamma(t) - \frac{E[I(C>t)\bar{G}(t|\mathbf{V})^{-1}\gamma(t)\bar{G}\{t|\bar{\mathbf{W}}(t-)\}Y(t)\exp\{\beta\gamma(t)\}]}{E[I(C>t)\bar{G}(t|\mathbf{V})^{-1}\bar{G}\{t|\bar{\mathbf{W}}(t-)\}Y(t)\exp\{\beta\gamma(t)\}]}\right) \times \frac{\bar{G}\{t|\bar{\mathbf{W}}(t-)\}\tilde{\Delta}^*(t)dM_{\beta,\lambda_0}(t)}{\bar{G}(t|\mathbf{V})}$$

where $D_h^*(\mathbf{V}|\beta,\lambda_0) = \int_0^\tau h^*\{t,\bar{\mathbf{W}}(t-)\}dM_{\beta,\lambda_0}(t)$. The estimator for the baseline hazard $\lambda_0(t)$ given an estimator $\hat{\bar{G}}$ of $\bar{G}$ is

$$\hat{\lambda}_0(t|\beta) = \sum_{i=1}^n \frac{\tilde{\Delta}_i(t)\hat{\bar{G}}\{t|\bar{\mathbf{W}}_i(t)\}}{\hat{\bar{G}}(t|\mathbf{V}_i)}dN_i(t) \Big/ \sum_{i=1}^n \frac{\tilde{\Delta}_i(t)\hat{\bar{G}}\{t|\bar{\mathbf{W}}_i(t)\}}{\hat{\bar{G}}(t|\mathbf{V}_i)}Y_i(t)\exp\{\beta\gamma_i(t)\}$$

Given estimators $\hat{h}^*$, $\hat{\bar{G}}$ and $\hat{\lambda}_0$ of h^*, $\bar{G}$ and λ_0, an estimator for β can be obtained by solving the estimating equation:

$$\sum_{i=1}^n \mathbf{U}_G\{\mathbf{E}_i|\hat{\bar{G}},\hat{D}_h^*(.|\beta,\hat{\lambda}_0)\} = \mathbf{0}$$

Note that $\bar{G}$ is estimated by fitting the multiplicative intensity model for the censoring process. The estimate for h^* is then obtained by substituting $\hat{\bar{G}}$ for $\bar{G}$ and estimating the expectations empirically. The authors mention that one of the strengths of the method is that it can be easily implemented by adapting standard routines available in statistical software packages.

According to Miloslavsky et al (2004), the methods described above are readily applicable to the proportional rates model as well. In this case, consider $D_h^* = \int h^*\{t,\mathbf{Z}^*(t)\}dM_r(t)$ as a class of full data estimating functions, where $dM_r(t) \equiv dN(t) - E\{dN(t)|\mathbf{Z}^*(t-)\}$ and h^* is arbitrary. Thus, considering the same approach as

presented above, the authors defined the estimating equation

$$U_G^r(\mathbf{E}|D_h^*) = \int_0^\tau \left(\gamma^*(t) - \frac{E[I(C>t)\bar{G}(t|\mathbf{V})^{-1}\gamma^*(t)\bar{G}\{t|\mathbf{Z}^*(t-)\}Y(t)\exp\{\beta\gamma^*(t)\}]}{E[I(C>t)\bar{G}(t|\mathbf{V})^{-1}\bar{G}\{t|\mathbf{Z}^*(t-)\}Y(t)\exp\{\beta\gamma^*(t)\}]}\right) \times \frac{\bar{G}\{t|\mathbf{Z}^*(t-)\}\Delta(t)dM_r(t)}{\bar{G}(t|\mathbf{V})}.$$

Miloslavsky et al (2004) argues that these estimators are at least as efficient as the partial-likelihood-based estimating equations used in Lin et al (2000). These estimators remain consistent even if the censoring does not only depend on the covariates entering the proportional rates model as long as the censoring mechanism is estimated consistently and the identifiability assumption $P(C > \tau|V) > \delta > 0$ holds.

2.3 Varying Coefficient Models for Time-to-Event Data

2.3.1 Introduction

In many situations it is of interest to explore the functional form of the relationships between covariates and failure time and to examine whether the effects are changing over time. Various alternatives of the Cox model have been made to allow the coefficients to change over time. In such case, Murphy and Sen(1991) proposed a sieve estimation procedure, assuming that the coefficient functions are piecewise constants and Gamerman(1991) described a dynamic linear model approach by assuming that the baseline hazard and the coefficient functions are both piecewise constant functions. However, the piecewise constant assumptions from the former approaches may not be appropriate in some applications. A recent alternative approach was described by Cai and Sun(2003), who developed a local likelihood technique to estimate the time-dependent coefficients in Cox's regression model and provided asymptotic theory for the proposed

estimator.

Alternatively, several investigators (Sleeper and Harrington,1990; Gray, 1992; Hastie and Tibshirani, 1993; Nan et al, 2003) have used spline functions to model the relative risk in the Cox proportional hazards model. These approaches are generally used for testing hypothesis on covariates effects in a Cox-based model and also on how effects change over time in regression analysis of survival data. Hastie and Tibshirani(1993) used the smoothing spline penalized partial likelihood method to estimate the covariate effects in the Cox model. However, it is pointed out by other authors that the formal inference is not well developed in this setting, mainly due to the computation of the full information matrix, which is numerically very intensive (Gray, 1994; Cai and Sun, 2003). Other authors have proposed the use of additive models considering smoothing techniques (Sleeper and Harrington, 1990; Gray, 1992). Sleeper and Harrington(1990) used regression B-splines while Gray (1992) used a penalized smoothing B-spline approach. Nan et al(2003) considered similar approach as Sleeper and Harrington(1990). In the next sections we review with some details the approaches described by Gray (1992,1994) and by Nan et al(2003).

It is also worth to mention that all cited approaches were developed for settings of univariate time-to-event data. The only extension of using smoothing techniques for the multivariate failure time data was found in a recent paper by Berhane and Weissfeld (2003), which extended the Gray's model (1994) considering the marginal modelling setup of Wei, Lin and Weissfeld (1989). More details about this approach are presented in a later section.

2.3.2 B-Splines

B-splines, originally introduced by de Boor(1978), are a popular type of regression splines in statistical applications, mainly due to their numerical properties. To imple-

ment B-splines in survival analysis, the time axis is divided into $m+1$ intervals with T_{min} and T_{max}, respectively, denoting the beginning of follow-up and the latest time potentially observable in a given study (Giorgi et al, 2003).

Let t_1,..., t_m be the m interior knots with $T_{min} < t_1 < ... < t_m < T_{max}$. Let $t_{-(q-1)}, ..., t_{-1}$ and $t_{m+2}, ..., t_{m+q}$ be the $2(q-1)$ additional boundary knots, such that $t_{-(q-1)} = ... = t_{-1} = T_{min}$ and $t_{m+2} = ... = t_{m+q} = T_{max}$. In this setting, there are $m+q$ basis functions $B_{-(q-1),q}(t), ..., B_{m,q}(t)$ of order q (that is, degree $q-1$), which are recursively defined by:

$$B_{k,q}(t) = \frac{t - t_k}{t_{k+q-1} - t_k} B_{k,q-1}(t) + \frac{t_{k+q} - t}{t_{k+q} - t_{k+1}} B_{k+1,q-1}(t),$$

where $k = -(q-1), ..., m$, with $B_{k,1}(t) = 1$ if $t \in [t_k, t_{k+1})$ and $B_{k,1}(t) = 0$ otherwise. The basis functions $B_{k,q}$ are called B-splines.

Thus, the kth B-spline of order q is a weighted sum of the kth and (k+1)st B-spline of order *q-1*, with weights depending on the breakpoints and continuity conditions (Sleeper and Harrington, 1990). Therefore, each basis function is non-zero in a limited interval spanned by q adjacent knots which leads to stable estimates and reduces computations. Other important mathematical property of B-splines to be considered is that $\sum_{k=-(q-1)}^{m} B_{k,q}(t) = 1$, which shows that the basis is a partition of unity, and implies that any constant function lies in the span of the basis (Sleeper and Harrington, 1990).

The resulting B-spline function, $g(t)$, of order q with m interior knots, for $t \in (T_{min}, t_{m-1})$, is a linear combination of the basis functions

$$g(t) = \sum_{k=-(q-1)}^{m} \xi_k B_{k,q}(t)$$

Thus, conditional on the knots, the $B_{k,q}$ are known functions of t and implies the estimation of $q+m$ parameters. For instance, considering a cubic B-spline, we may write $g(t)$ as $\sum_{k=-3}^{m} \xi_k B_{k,4}(t)$. A commonly used modification of the cubic spline model is the natural cubic spline basis, which present a constraint to be linear beyond the two boundary knots (Ruppert, Wand and Carroll, 2003). Thus, the function $g(t)$ can be reparametrized using m+2 natural cubic B-splines basis functions, such that $g(t) = \sum_{k^*=1}^{m+2} \tilde{\xi}_{k^*} \ddot{B}_{k^*,4}(t)$. In the next sections we review approaches that consider the use of B-splines in the analysis of failure time data.

2.3.3 Estimation using B-Splines in Survival Analysis

2.3.3.1. Regression Splines

Many authors have discussed the use of regression splines for analysis of survival data (Sleeper and Harrington, 1990; Abrahamowicz, MacKenzie and Esdaile, 1996; Giorgi et al, 2003; Nan et al, 2003). In this section we review the varying coefficient Cox model presented by Nan et al (2003), which was proposed to investigate the association between two events, age at a specific bleeding pattern change and age at menopause, where both events were subject to censoring and their association varies with age at the marker event. The estimation proceeded using the regression spline method.
Suppose $W_i(t)$ is a time dependent variable and $\mathbf{Z}_i$ represents the baseline covariates. The varying-coefficient Cox model including both time-dependent and time-independent covariates is defined as:

$$\lambda_i(t|\mathbf{Z}_i, W_i(t)) = \lambda_0(t) \exp\{\beta' \mathbf{Z}_i + \theta(s) W_i(t)\}$$

where s represents the time of the occurrence of a marker event. In Nan's et al (2003) paper, for instance, $W_i = I(t \geq S_i)$, with S_i being the true age at the 60-day marker

event for woman i.

The estimation of the nonparametric function $\theta(s)$ is done through the regression spline method by approximating $\theta(s)$ using the natural cubic B-splines basis, where the natural spline was constraint to be linear beyond the two boundary knots (Nan et al, 2003).

The function $\theta(s)$ is parametrized using m+2 natural cubic B-spline basis functions $\ddot{B}_k(t),\ k = 1, ..., m+2$, such that $\theta(s) = \sum_{k=1}^{m+2} \tilde{\xi}_k \ddot{B}_s(t)$. Thus, replacing $\theta(s)$ by its B-spline approximation in the previous varying-coefficient model, we have:

$$\lambda_i(t|\mathbf{Z}_i, W_i(t)) = \lambda_0(t) \exp\{\beta' \mathbf{Z}_i + \tilde{\xi}' \tilde{\mathbf{W}}_i(t)\},$$

where $\tilde{\xi} = (\tilde{\xi}_1, ..., \tilde{\xi}_{m+2})'$ and $\tilde{\mathbf{W}}_i(t) = (\ddot{B}_1(s)W_i(t), ..., \ddot{B}_{m+2}(s)W_i(t))'$.

The estimation of the parameter vectors $(\beta, \tilde{\xi})$ is then obtained by maximizing the following log partial likelihood function:

$$\ell(\beta, \tilde{\xi}) = \sum_{i=1}^{n} \int \left[\beta' \mathbf{Z}_i + \tilde{\xi}' \tilde{\mathbf{W}}_i(t) - \log \sum_{j=1}^{n} I(X_j \geq t) \exp\{\beta' \mathbf{Z}_j + \tilde{\xi}' \tilde{\mathbf{W}}_j(t)\} \right] dN_i(t),$$

where $X_i = min(C_i, T_i)$. Because the spline fit is just Cox regression on constructed regressors, the maximum partial likelihood estimators of β and $\tilde{\xi}$ can be obtained by any statistical package that implements Cox regression (Nan et al, 2003). Therefore, the nonparametric function $\theta(s)$ can be estimated by $\hat{\theta}(s) = \sum_{k=1}^{m+2} \hat{\tilde{\xi}}_k \ddot{B}_k(s)$. The pointwise confidence interval for $\hat{\theta}(s)$ can be estimated using its variance estimator $var\{\hat{\theta}(s)\} = \ddot{\mathbf{B}}'(s) cov(\hat{\tilde{\xi}}) \ddot{\mathbf{B}}(s)$.

Estimation of the baseline cumulative hazard function $\Lambda_0(t)$, given the occurrence

of a marker event at S_i, is done by using Breslow estimator:

$$\hat{\Lambda}_0(t) = \int_0^t \left[\sum_{i=1}^{n} I(X_i \geq u) \exp\{\hat{\theta}(S_i) W_i(u) + \hat{\beta}' \mathbf{Z}_i\} \right]^{-1} \left\{ \sum_{i=1}^{n} dN_i(u) \right\},$$

2.3.3.1.1. Estimation of Number of Knots

Even though the regression splines present advantages that the models can be fitted using standard software and inferences are made using standard techniques, one of the major drawbacks of such approach is that they are also very sensitive to the number and location of knots (Gray, 1992; Therneau and Grambsch, 2001). Therefore, in order to define the optimal number and location of knots, many criteria have been defined, including cross-validation (CV) criterion, generalized cross-validation(GCV), Mallow's C_p criterion and Akaike's information criterion (AIC) (Ruppert, Wand and Carroll, 2003).

In their paper, Nan et al (2003) extended the cross-validation (CV) and generalized cross-validation (GCV) methods proposed by O'Sullivan (1988) to choose the number of knots in the regression spline setting, taking into account that $\Lambda_0(t)$ is unknown and is estimated.

In order to ensure an approximate equal number of events in each interval, the location of interior knots is usually based on the quantiles of the distribution of the observed event times (Giorgi et al., 2003).

Simulation studies showed that the pointwise biases of the B-spline estimator $\hat{\theta}(.)$ are close to zero, and the pointwise model based SEs of $\hat{\theta}(.)$ agree well with their empirical counterparts, except for the boundary (Nan et al, 2003).

2.3.3.2. Penalized B-Splines

An alternative to the regression spline approach is a penalized/smoothing spline.

Consider the following time-varying coefficient model:

$$\lambda_i(t|\mathbf{Z}_i, W_i) = \lambda_0(t)\exp(\beta'\mathbf{Z}_i + \theta(t)W_i),\ t \geq 0$$

This model represents non-proportional hazards unless $\theta(t)$ is a constant. Considering the simple case where there is only one covariate W, such that W is an indicator for a group (W=0 or W=1) then $\theta(t)$ measures the difference in log(relative risk) between the two groups (Hastie and Tibshirani,1993). The covariates can be either fixed or time-varying.

For estimating the parameters from the previous model, Gray (1992) proposed a penalized B-spline based model, in which $\theta(t)$ is substituted with the flexible form $g(t)$:

$$\lambda_i(t|\mathbf{Z}_i, W_i) = \lambda_0(t)\exp(\beta'\mathbf{Z}_i + g(t)W_i),\ t \geq 0$$

where the parametrization of $g(t)$ is given by $g(t) = \sum_{k=-(q-1)}^{m} \xi_k B_{k,q}(t)$, $t \in (T_{min}, t_{m-1})$.

Thus, the model can be rewritten as

$$\lambda_i(t|\mathbf{Z}_i, W_i) = \lambda_0(t)\exp\left(\beta'\mathbf{Z}_i + \sum_{k=-(q-1)}^{m} \xi_k B_{k,q}(t)W_i\right),\ t \geq 0$$

where $B_k's$ may be considered to be (m+4) standard cubic B-spline basis functions (e.g., q=4), as defined in section 2.3.2, with the number and location of knots as defined in the next subsection. For simplicity, let's drop the index q from the B-spline notation. Thus, a cubic spline basis with m interior knots would have m+4 functions $B_1(t), ..., B_{m+4}(t)$. Considering the B-spline property $\sum_{k=-(q-1)}^{m} B_{k,q}(t) = 1$, $g(t)$ can be reparametrized

and written in terms of $m+3$ spline terms. Thus, $g(t)$ can be rewritten as

$$g(t) = \gamma_0 + \sum_{k=1}^{m+3} \gamma_k B_k(t)$$

The penalty considered in the estimation procedure is given by the integral of a squared higher derivative of the fitted curve (Eilers and Marx, 1996), such that the standard form of the penalty function for cubic splines is given as:

$$\frac{1}{2}\alpha \int_{min(t_i)}^{max(t_i)} [g''(t)]^2 dt,$$

where α controls the amount of smoothing applied, with $\alpha = 0$ and $\alpha = \infty$ leading to a non-penalty regression spline function and a linear term, respectively (Berhane and Weissfeld, 2003). Theoretically, it is assumed that the amount of smoothing is fixed and specified a priori. Because the penalty function is a quadratic function of $\tilde{\gamma}' = (\gamma_1, ..., \gamma_{m+3})$, it can be rewritten as $\frac{1}{2}\alpha\tilde{\gamma}'\tilde{\mathbf{D}}\tilde{\gamma}$, where $\tilde{\mathbf{D}}$ is an appropriately chosen symmetric, nonnegative matrix, such that the i,j th element of $\tilde{\mathbf{D}}$ would be $\int_{min(t_i)}^{max(t_i)} B_i''(t)B_j''(t)dt$.

Hence, the parameter estimates are obtained by maximizing the log penalized partial likelihood, which is defined as

$$\ell_p(\beta, \gamma) = \ell(\beta, \gamma) - \frac{1}{2}\alpha\gamma'\mathbf{D}\gamma,$$

where $\gamma' = (\gamma_0, \tilde{\gamma}')$. For the cubic splines setting, $\mathbf{D}$ is an $(m+4) \times (m+4)$ matrix with the first row and column being zeros, since the constant term passes unpenalized.

The unpenalized partial likelihood can be written as

$$L(\beta,\gamma)=\prod_{i=1}^{n}\left(\frac{\exp(\gamma_0(t)+\sum_{k=1}^{m+2}\gamma_k B_k(t)W_i+\beta\mathbf{Z}_i)}{\sum_{l\in\Re}\exp(\gamma_0(t)+\sum_{k=1}^{m+2}\gamma_k B_k(t)W_i+\beta\mathbf{Z}_i)}\right)^{\Delta_i}$$

Instead of using the standard form of the penalty function for cubic splines as $\frac{1}{2}\alpha\int[g''(t)]^2dt$ because the cubic splines tend to be unstable in the right tail of the distribution as discussed by Gray (1992), quadratic splines may be used with penalty $\frac{1}{2}\alpha\int[g'(t)]^2dt$. Alternatively, Gray (1992) also recommends, for computational reasons, the use of piecewise constant functions with penalty $\frac{1}{2}\alpha\sum_{k=1}^{m+2}(\gamma_k-\gamma_{k-1})^2dt$.

In any case, let $\eta=(\beta,\gamma)$. Thus, it can be shown that $\sqrt{n}(\hat{\eta}-\eta_{(T)})=n(\mathcal{I}+\alpha\bar{\mathbf{D}})^{-1}n^{-\frac{1}{2}}\mathbf{U}(\gamma_{(T)})+\circ_p(1)$, where $\bar{\mathbf{D}}$ is the expanded penalty matrix that augments rows and columns of zeros to $\tilde{\mathbf{D}}$ to account for the unpenalized terms in the model; $\mathbf{U}(\gamma_{(T)})$ is the unpenalized score vector; $\mathcal{I}$ is the information matrix and $\eta_{(T)}$ is the true parameter value (Gray, 1994).

Then it follows from the asymptotic normality of $\mathbf{U}(\gamma_{(T)})$ that $\sqrt{n}(\hat{\eta}-\eta_{(T)})$ is asymptotically normal with mean $\mathbf{0}$ and variance given as the limit of $n\mathbf{V}$, where $\mathbf{V}=(\mathcal{I}+\alpha\bar{\mathbf{D}})^{-1}\mathcal{I}(\mathcal{I}+\alpha\bar{\mathbf{D}})^{-1}$. These asymptotic results assume that the number of terms in the spline functions is held fixed as $n\to\infty$.

By analogy with the usual (unpenalized) parametric likelihood procedures, Gray (1994) proposed three different test statistics considering the null hypothesis $\gamma=\mathbf{0}$:

- A penalized Wald-type statistics can be defined as $Q_w=\hat{\gamma}'(\mathcal{I}_{\gamma\gamma|\beta}+\alpha\mathbf{D})\hat{\gamma}$, where $\mathcal{I}_{\gamma\gamma|\beta}=\mathcal{I}_{\gamma\gamma}-\mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\gamma}$, with $\mathcal{I}_{\gamma\gamma}$ denoting the derivatives with respect to γ;

- A penalized quadratic score statistic is defined by $Q_s=\mathbf{S}'_{\gamma}(\hat{\beta}_0,\mathbf{0})(\mathcal{I}_{\gamma\gamma|\beta}+\alpha\mathbf{D})^{-1}\mathbf{S}_{\gamma}(\hat{\beta}_0,\mathbf{0})$, where $\mathbf{S}(\beta,\gamma)=(\mathbf{S}'_{\beta}(\beta,\gamma),\mathbf{S}'_{\gamma}(\beta,\gamma)'$, $\hat{\beta}_0$ is the maximum partial likelihood estimator for β when $\gamma=\mathbf{0}$, and $\mathbf{S}_{\gamma}(\hat{\beta}_0,0)=\partial L_p(\hat{\beta}_0,0)/\partial\gamma$;

- A penalized likelihood ratio statistic is given by $Q_l = 2[L_p(\hat{\beta}, \hat{\gamma}) - L_p(\hat{\beta}_0, 0)]$.

Under the null hypothesis, the statistics Q_s, Q_l and Q_w all have the same asymptotic distribution, which is that of $\sum e_j Z_j^2$, where Z_j are independent standard normal random variables and the e_j are the eigenvalues of the matrix $\lim \mathcal{I}_{\gamma\gamma|\beta}(\mathcal{I}_{\gamma\gamma|\beta} + \alpha\mathbf{D})^{-1}$. Thus, the reference distribution for the test statistics under H_0 is given by a weighted sum of independent χ_1^2's, where the weights are given by the eigenvalues defined previously. One of the advantages of this approach over nonparametric smoothing splines is that here e_j are easy to compute (Gray, 1994).

2.3.3.2.1. Choice of Smoothing Parameter, Degrees of Freedom and Placement of Knots

Based on numerical results, Gray (1994) suggested that the number and location of the knots is not very important as long as the number is large enough (between 10 and 15) and they are reasonably spread out. The general algorithm used by this author is to put roughly equal amounts of data between the knots.

Theoretically the smoothing parameter is considered fixed and defined a priori in order to be used for obtaining the parameter estimates. However, operationally the smoothing parameter is calculated by the following relationship with the degrees of freedom, which should be specified for each nonparametric term, such that:

$$df = trace\{\lim \mathcal{I}_{\gamma\gamma|\eta}(\mathcal{I}_{\gamma\gamma|\eta} + \alpha\mathbf{D})^{-1}\}$$

Thus, the above expression for the values of the smoothing parameter that give the specified degrees of freedom should be solved (Gray, 1992). Based on this definition, Berhane and Weissfeld (2003) suggest that the number of knots that determine the B-spline basis functions are set to be at least twice the number of degrees of freedom.

Some authors mention that even though it is possible to develop some automatic procedure to select smoothing parameters by using criteria such as GCV, its implications for hypothesis testing are not obvious (Gray, 1994; Berhane and Weissfeld, 2003).

2.3.4 Extensions for Multivariate Time-to-Event Data

An extension of Gray's model for multivariate time-to-event data was proposed by Berhane and Weissfeld (2003), whose focus is on the ability to conduct simultaneous inference on several time-to-event outcomes in models in which the exposure-response relationships may have nonlinear forms. The authors considered a study with G different time-to-event outcomes, such that the proportional hazards model for the gth outcome is of the form

$$\lambda_{gi}(t) = \lambda_{g0}(t) \exp\left\{ \sum_j \beta_{jg}{}' \mathbf{Z}_{jgi} + f_g(h_{gi}) \right\}, \; t \geq 0$$

where h_g is a non-parametric function of one additional covariate. The penalized regression spline approach is used to estimate $f_g(h_g) = \gamma_{0g} h_g + \sum_{k=1}^{m+2} \gamma_{kg} B^*_{kg}(h_g)$. The inference for each of the marginal models can be done using the developments found in Gray (1994). Thus, the authors extended the methods described in Wei et al.(1989) to test for trends across parameter estimates and to combine estimates across margins to test for covariate effects of interest.

In order to take into account that $\eta_g = (\beta_g, \gamma_g)$ across G multiple outcomes are generally correlated, the authors used analogous developments as in Wei et al (1989) to show that the asymptotic covariance matrix between $\sqrt{n}(\hat{\eta}_g - \eta_g)$ and $\sqrt{n}(\hat{\eta}_v - \eta_v)$ can be consistently estimated. Berhane and Weissfeld(2003) also discuss methods for conducting simultaneous inference of the overall effect and/or linearity of the nonparametric term, h, across failure types, using extension of the results presented in Gray

(1994). However, there is no discussion about considering time-varying coefficients in their model.

2.4 Overview of Proposed Research Work

So far, almost all methods proposed in the literature for the analysis of time-varying coefficients in survival analysis have been developed for univariate time-to-event data. To our knowledge, the approach for the estimation of varying-coefficients considering B-splines in the WLW model is the only one proposed for multivariate time-to-event data settings (Berhane and Weissfeld, 2003). An appealing approach for the analysis of recurrent event data is the marginal means/rates model, which we extended by incorporating B-splines for the estimation of time-varying effects. Thus, we consider the inclusion of time-varying effects in the marginal rates models in this research and we discuss two estimation approaches for these time-varying coefficients: regression and penalized B-splines. The proposed methods are described in details in Chapters 3 and 4, respectively. We also extended the generalized cross-validation (GCV) criteria for the estimation of number of knots in the context of our proposed regression B-spline model.

In addition, we conducted extensive simulation studies to compare two recently proposed methods by Wang et al(2001) and Miloslavsky et al(2004) for analyzing recurrent time-to-event data in the presence of dependent censoring. The results of this empirical comparison are presented in Chapter 5.

To illustrate the proposed methods, we apply both the traditional and proposed methods to the aforementioned vitamin A randomized clinical trial. Using this data, we evaluate how the effect of supplementation of vitamin A on diarrhea morbidity behaves over time, taking also into account the effect of other factors, such as child's

age and gender, for instance. In the following Section, we provide a detailed description of the vitamin A clinical trial as well as some preliminary results of applying the most commonly used methods for recurrent event data.

2.4.1 Motivating Example: Vitamin A Community Trial

Diarrhea is one of the most prevalent causes of child mortality worldwide, killing over 1 million children annually (Victora et al., 2000). Tragically, methods of prevention, treatment and management are well understood but not available to those who need it most. Major causes of diarrhea include limited access to safe drinking water, fecal contamination of surroundings, and unsanitary conditions. Untreated, diarrhea can quickly cause mortality by dehydration. Children dehydrate more quickly than adults and thus are more vulnerable to diarrheal mortality. Diarrhea is ultimately fatal to approximately one in every 200 children who contract it (UNICEF, 2003). As a chronic condition, diarrhea compromises the integrity of the entire body, reducing immune system capabilities, growth, development, and general nutritional and energy stores.

Viruses, bacteria and protozoa may be responsible for diarrheal illnesses. Exposure to some pathogens can vary with the season. The mechanism of transmission of diarrhea may be through foodborne transmission (a common pathogen would be E.coli), waterborne transmission (pathogens include Giardia lambia and Shigella sp.) and transmission by direct contact. In settings where it is difficult to maintain good hygiene, person-to-person transmission is very common and affects particularly small children. In this case, rotavirus is an important cause of diarrhea in pediatric populations, seen predominantly in infants, and Giardia lambia, in toddlers. A particular child can experience repeated episodes of diarrhea, which can be related for at least two reasons: the prior infection made the person less resistant to a subsequent exposure; or the factors that led to the first exposure led also for the second exposure (Byers,

Guerrant and Farr, 2001).

The prevention of diarrhea involves commitment on the part of both individuals and communities. On an individual level, exclusive breastfeeding for infants is the first line of defense and is highly recommended on many other levels, as it is inexpensive and unparalleled in benefit to infant and mother (Bhandari et al., 2003; Morris et al., 1999). Vigilance in maintaining personal and domestic hygiene behaviors is another way to control the risks latent in the immediate environment of children. On a community level, mobilization of sanitary water supplies can not only prevent diarrhea by eliminating risk due to contaminated water; an ample availability of sanitary water can also facilitate individual and community efforts to maintain hygienic personal and domestic behaviors. Comprehensive vaccination against infectious diseases will minimize weakening of GI tract flora and render children less susceptible to diarrhea (WHO, 1999). Vitamin A deficiency is other factor that may affects the immune system, reducing the immune response to children's infections.

Scientific information accumulated in the last decades has led to consensus about the effect of vitamin A supplementation on the reduction of child mortality by 23 to 34 % in populations where vitamin A deficiency is endemic, averting up to one million deaths a year. This reduction was due in large part to a fall in diarrheal and measles-related deaths in the supplemented children. The role of vitamin A supplementation on diarrheal morbidity however is less clear than on mortality.

In the last two decades, many clinical trials had been conducted to evaluate the effect of the supplementation of vitamin A on diarrhea morbidity and other related illnesses in areas where its intake is inadequate. One of such studies was a randomized community trial 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. 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. For children aged 1-4 years, WHO recommends that the optimal interval between doses is 4 to 6 months and that the minimum interval doses for the prevention of vitamin A deficiency is one month.

The guidelines proposed by WHO as well as by other available sources do not provide specific information on how long a high-dose vitamin A supplementation affects diarrhea morbidity. As diarrhea is still a major cause of morbidity and mortality in small children in developing countries, it would be of interest to evaluate more comprehensively the effect of vitamin A supplementation on diarrheal episodes in children aged 6 months to 5 years. Thus we consider the data from the Brazilian vitamin A clinical trial in order to further evaluate the research questions of interest about the effect of vitamin A supplementation on diarrhea morbidity.

For the Brazilian study, 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. A complete investigation of signs and symptoms was conducted when diarrhea was reported. 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 at the household, energy and water supply, among others. 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. More details about this study is presented in Barreto et al (1994).

Since we are interested in evaluating the effect of vitamin A supplementation on the occurrence of recurrent diarrheal episodes, it's worth to present some important

definitions in this setting as follows: *(i)* As the household visits collected information regarding to each 24 hour period, the study defines **24 h period** as the time from the moment the child woke up one day until he/she woke up the next day.*(ii)* **A day with diarrhea** was defined as 3 or more liquid or semi-liquid motions reported in a 24 h period. The number of motions per 24h was recorded. Thus, *(iii)***an episode of diarrhea** was defined as a sequence of days with diarrhea and the episode was considered finished when there were 3 or more days without diarrhea. *(iv)*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 h period. Thus, we defined three groups: mild (duration ≤ 2), moderate (duration ≥ 3 and average number of motions < 5) and severe (duration ≥ 3 and average number of motions ≥ 5) episodes.

2.4.1.1. Description of the Data

The analyses presented here include 1,207 children with mean age 27.3 months at baseline (std dev=12.1, range=6-48 months), being 52.4 % boys and 50.1% randomized to receive vitamin A (treatment group). 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 episodes of diarrhea by 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. Table 2.1 presents the distribution of the number of episodes by treatment group. According to the data presented in Table 2.1, 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. However, no statistically significant difference was found between the number

of episodes of diarrhea by treatment group ($\chi^2 = 1.28$, p=0.8649).

TABLE 2.1: Distribution of number of diarrheal episodes by treatment group

Number of episodes	Vitamin A	Placebo
0-3	271 (44.79%)	256 (42.52%)
4-7	157 (25.95%)	153 (25.42%)
8-11	89 (14.71%)	94 (15.61%)
12-15	49 (8.10%)	54 (8.97%)
≥ 16	39 (6.45%)	45 (7.48%)
Total	605	602

The overall number of episodes of diarrhea to be consider in this project 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. Only 276 (3.88 %) of the episodes were considered severe. The distribution of episodes by severity level and treatment group is presented in Table 2.2. Note that we concatenated the mild and moderate groups in this table. Using GEE, we found some evidence that the proportion of severe episodes in vitamin A (3.3%) and placebo (4.5%) may be different ($\chi^2 = 2.78$, p-value=0.0957).

TABLE 2.2: Distribution of diarrheal episodes by severity and treatment group

Severity	Vitamin A	Placebo
Mild-Moderate	3351(96.74%)	3482(95.53%)
Severe	113(3.26%)	163(4.47%)

The distribution of the number of episodes in each of the intervals between the four dosages of vitamin A is presented in Table 2.3. According to these results, 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% (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). Same trend was observed in both treatment groups.

TABLE 2.3: Distribution of diarrheal episodes by interval of occurrence and treatment group

Interval	Vitamin A	Placebo
1^{st} dosage cycle	1441(41.60%)	1606(44.06%)
2^{nd} dosage cycle	1294(37.36%)	1336(36.65%)
3^{rd} dosage cycle	729(21.05%)	703(19.29%)

2.4.1.2. Application of methods for analysis of recurrent time-to-event data

The most commonly used models for recurrent time-to-event data were applied to this data to evaluate the effect of vitamin A supplementation on childhood morbidity. The event outcome was diarrhea. For these analyses, the i^{th} child contributes $K_i + 1$ records, where K_i represents the number of observed events for the i^{th} child. Using counting process notation, we have that for the k^{th} record of the i^{th} child, *tstart* is the time of the (k-1)th event (or 0 if k=1) and *tstop* is the time of the kth event (or censoring time if $k = K_i + 1$). 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.

Initially, we present results related to the analysis of data regarding the first treatment cycle, i.e., data from the day of receiving the first treatment dose until the day just before the second dose, which was approximately four months after the first dose. We are defining the total time as the time from the first dose of vitamin A until the occurrence of an episode of diarrhea. The only covariate considered is treatment, which is 0 if the child received placebo and 1 if the child received vitamin A. During this period we observed 3,047 episodes of diarrhea.

The results from the AG model and Marginal Rates/Means model considering all episodes during this first interval (i.e.,between first and second treatment doses) are displayed in Table 2.4. Based on AG model, the results show that vitamin A supplementation has a significant effect on diarrheal episodes ($\hat{\beta}$=-0.12; 95% CI=(-0.21; -0.03)). That is, the hazard of experiencing an episode of diarrhea since first dosage is 11.3% lower for those who received vitamin A compared to those who received placebo. The marginal means model also showed a significant negative association between diarrhea and treatment.

TABLE 2.4: Treatment estimates for the Vitamin A trial during first dosage cycle

Model	$\hat{\beta}$	Estimated robust SE($\hat{\beta}$)	$\{\hat{\beta}/SE(\hat{\beta})\}^2$
Andersen-Gill	-0.120	0.0444	7.323
Marginal Means	-0.120	0.0590	4.144

During the first interval (between first two dosages), only 74(12.23%) and 86(14.29%) of the children in the vitamin A and placebo groups, respectively, had more than five episodes. Therefore, the event-specific estimates may be unreliable for greater than five episodes. Hence, to allow direct comparison between all models the data were truncated after five events. The results for all fitted models are shown in Table 2.5. The patterns of the results for the models are remarkably similar, especially regarding the estimation of the common effect using the PWP-Total time (PWP-TT: $\hat{\beta}$=-0.112), AG and the marginal means models ($\hat{\beta}$=-0.116). The common estimate for WLW model is obtained by using optimal weights for estimating the average treatment effect across the marginal models. The overall treatment estimate for the WLW model ($\hat{\beta}$=-0.144) was higher than those obtained by the previous models. These results seem to be consistent with the carry-over effect of the WLW method described by some authors (Kelly and Lim, 2000) since the overall estimate reflects the weighted average of the

event-specific estimates. The estimate from the LWA model also seems to overestimate the treatment effect ($\hat{\beta}$=-0.154), which may be due to the fact that the LWA method allows subjects to be at risk several times for the same event. It is worth to mention that, considering any of those methods, the overall treatment effect was found to be statistically significant.

Regarding the event-specific estimates, the association between treatment effect and time to first episode of diarrhea was not significant ($\hat{\beta}$=-0.127, SE($\hat{\beta}$)=0.0658). The treatment effect seems to be significantly associated only to the time to second episode. Hence, it seems that the treatment effect changes as the number of events increases. The PWP gap time model (PWP-GT) shows similar patterns. However, the interpretation of event-specific estimates is not very simple. For the conditional models (PWP-TT and PWP-GT), the interpretation of the estimated parameters are conditional on having had previous events because the analysis is based on restricted risk sets. Thus, the decrease of the importance of the treatment as the number of events increases may be related to the fact that the children with a higher number of episodes are progressively less healthy, which was one of the hypothesis mentioned by Cai and Schaubel (2004).

The WLW method, on the other hand, has been criticized for analysis of recurrent time-to-event data because it is based on unrestricted risk sets, in which the subjects can be at risk for the (k+1)th event prior to having experienced the kth event. Our results indicate that the magnitude of the estimates of the WLW model for all subsequent events are higher than those from PWP-TT, even though they are still not statistically significant. Therefore, this may illustrates the carry-over phenomenon noted in Kelly and Lim (2000) and Cai and Schaubel (2004). This happens especially when a covariate effect exists for the first events, but not for subsequent recurrences, resulting in larger estimates for those latter events due to the carry-over effect.

TABLE 2.5: Treatment estimates, considering first five diarrheal episodes, for the Vitamin A trial during first dosage cycle

Model	Estimates	$\hat{\beta}$	Estimated robust SE($\hat{\beta}$)	$\{\hat{\beta}/SE(\hat{\beta})\}^2$
WLW	Episode 1	-0.127	0.0658	3.765
	Episode 2	-0.208	0.0758	7.521
	Episode 3	-0.177	0.0904	3.830
	Episode 4	-0.168	0.1075	2.441
	Episode 5	-0.242	0.1310	3.415
	Common	-0.144	0.0640	5.056
LWA	Common	-0.154	0.0631	5.984
PWP-TT	Episode 1	-0.127	0.0658	3.701
	Episode 2	-0.156	0.0761	4.205
	Episode 3	-0.026	0.0907	0.084
	Episode 4	-0.062	0.1079	0.331
	Episode 5	-0.176	0.1315	1.797
	Common	-0.112	0.0387	8.392
PWP-GT	Episode 1	-0.127	0.0658	3.701
	Episode 2	-0.153	0.0761	4.045
	Episode 3	-0.011	0.0907	0.014
	Episode 4	-0.022	0.1081	0.043
	Episode 5	-0.045	0.1320	0.011
	Common	-0.092	0.0387	5.646
Andersen-Gill	Common	-0.116	0.0445	8.968
Marginal Means	Common	-0.116	0.0499	5.361

A rate model with the effect being piecewise constant was implemented for a further evaluation of the treatment effect over the period between the first and second dosage in the vitamin A trial. Thus, we defined indicators considering different time intervals, which allows us to evaluate whether the effect of vitamin A supplementation varies with time. Table 2.6 displays the results for the evaluation of treatment effect in intervals of 15 and 60 days, respectively. For the model using intervals of 15 days, we had to

put the two last intervals together because there were only very few events on the last interval (135-150 days), resulting in unstable estimates.

TABLE 2.6: Treatment estimates, considering different time intervals, for the Vitamin A trial during first dosage cycle

Model	Intervals	$\hat{\beta}$	Estimated SE($\hat{\beta}$)	p-value
Treatment	0-15	-0.1365	0.1477	0.3552
	15-30	-0.0591	0.1016	0.5608
	30-45	-0.2375	0.1110	0.0324
	45-60	-0.2447	0.1060	0.0210
	60-75	0.0109	0.1084	0.9196
	75-90	-0.2442	0.1208	0.0432
	90-105	-0.1583	0.1143	0.1661
	105-120	0.0414	0.1062	0.6970
	120-150	-0.0673	0.1289	0.6014
Treatment	0-60	-0.1716	0.0694	0.0134
	60-120	-0.0777	0.0724	0.2831
	120-150	-0.0673	0.1288	0.6015

The results show some evidence that the effect of treatment may vary somewhat with time on study (Table 2.6). Both models suggest that the benefit from vitamin A supplementation appears to be smaller or disappearing later on in the study. In order to better visualizing how treatment effect changes over time, Figure 2.1 shows the estimates for treatment effect using intervals of 15 days along with a smoothing curve. In situations such that, when a treatment may gradually lose effectiveness or a treatment may have a latent period of minimal effectiveness before the required cumulative dose is attained, one alternative approach is to consider models with time-dependent coefficients, $\theta(t)$. When $\theta(t)$ is not constant over time, the impact of one or more covariates on the hazard may vary over time (Therneau and Grambsch, 2001).

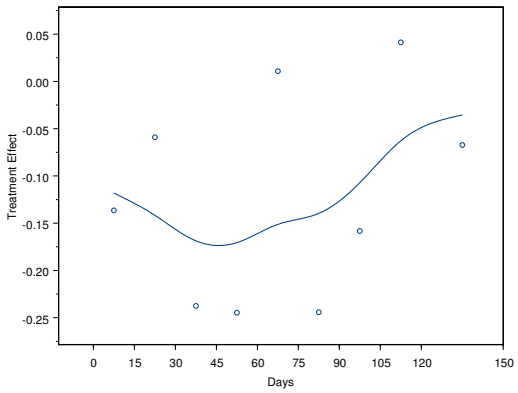

FIGURE 2.1: Treatment Effect using 15 days interval for the Vitamin A trial

In Chapters 3 and 4 we reanalyze this data using the proposed models as well as we consider a more comprehensive evaluation of the effect of vitamin A supplementation on diarrhea morbidity, taking into account the multiple dosage scheme that was used in the Vitamin A clinical trial.

CHAPTER 3

REGRESSION SPLINES IN THE TIME-VARYING COEFFICIENT RATES MODEL

3.1 Introduction

Many epidemiologic studies involve the occurrence of recurrent events, such as times to opportunistic infections among AIDS patients or to lung exacerbations in cystic fibrosis patients, and much attention has been given for the development of modelling techniques that take into account the dependence structure of multiple event data in the last few decades (Prentice, Williams and Peterson, 1981; Andersen and Gill, 1982; Wei, Lin and Weissfeld,1989; Pepe and Cai, 1993; Lin et al., 2000). Recent papers (Kelly and Lim, 2000; Cai and Schaubel, 2004) have discussed the appropriateness of such approaches to handle recurrent event data.

Recent research has been focusing on more complex recurrent event settings which include large number of recurrent events, time-dependent covariates, time-dependent effects and dependent censoring among other features (Wang, Qin and Chiang, 2001; Duchateau et al., 2003; Ghosh and Lin, 2003; Miloslavsky et al., 2004). Much effort has also been devoted to the development of methods for the estimation of means/rates of recurrent events in recent years (Pepe and Cai, 1993; Lawless and Nadeau, 1995; Lin et al, 2000). The main appeal of using such approaches is that the mean number of events

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

In many situations it is also 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 (Hastie and Tibshirani, 1990; Sleeper and Harrington, 1990; Gray, 1992; Nan et al., 2003) are all Cox-based models defined for univariate time-to-event. To our knowledge, no time-varying coefficient model had been proposed to handle any type of multiple time-to-event outcomes. Thus, in this Chapter we propose a method for estimating time-varying coefficients in the marginal rate models using regression B-splines.

We illustrate the application 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 pre-school age children, who were assigned to receive either placebo or vitamin A every 4 months for one year . 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 primarily 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 (Barreto et al, 1994). 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 (1997), use of time-to-event models will lead to greater efficiency and accuracy. Rate models have been used to analyze 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.

Preliminary analysis of the vitamin A study suggested that the effect of vitamin A supplementation on recurrent diarrhea may change over time. Therefore, it is important

to develop methods to estimate such time-varying effects. Hence, the main purpose of this Chapter is to present a statistical method that incorporates B-splines for estimation of time-varying coefficients in modeling recurrent time-to-event data. The remainder of this Chapter is organized as follows. In Section 2, we describe the model of interest and present the proposed methods. Simulation methods and results are discussed in Section 3. In Section 4, the model is applied to the vitamin A community trial data that was described in details in Chapter 2. A discussion of issues pertinent to the proposed method and its application is given in Section 5.

3.2 Model and Methods

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 marginal rates model that incorporates a smoothing technique called regression splines. The marginal rate model may be written:

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

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 $\mathbf{W}(t)$ 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 the time-dependent effect $\theta(t)$ might be done by using standard cubic B-spline basis functions $\tilde{B}_k(t)$,$(k = 1, ..., m+3)$, such that $\theta(t) = \gamma_0 + \sum_{k=1}^{m+3} \gamma_k \tilde{B}_k(t)$. B-splines, originally introduced by de Boor(1978), are popular types of regression splines in statistical applications, mainly due to their numerical properties. The proposed model uses products of a covariate and 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 $\gamma = (\gamma_0, \gamma_1, ..., \gamma_{m+3})'$; $\tilde{\mathbf{W}}_i(t) = (W_i(t), \tilde{B}_1(t)W_i(t), ..., \tilde{B}_{m+3}(t)W_i(t))'$.

Note that the model above does not include time-dependent coefficients. Now the model becomes a standard marginal rates model with time-dependent covariate $\tilde{\mathbf{W}}(t)$. Hence, the log of the partial likelihood is given by:

$$\ell = \sum_{i=1}^{n} \int_0^\tau \left[(\beta' \mathbf{Z}_i(t) + \gamma' \tilde{\mathbf{W}}_\mathbf{i}(\mathbf{u})) - \log(\sum_{j=1}^{n} Y_j(u) \exp\{\beta' \mathbf{Z}_j(t) + {}'\gamma' \tilde{\mathbf{W}}_\mathbf{j}(u)\}) \right] dN_i(u),$$

where $dN_i(u)$ denotes the number of events in a small time interval $[u, u + du]$, $Y_i(t) = I(C_i \geq t)$ is the at-risk indicator and τ denotes the end of the follow-up period.

Thus, considering the regularity conditions for the marginal rates model (Lin et al, 2001), the estimates of the regression parameters are obtained by the solution of the following unbiased estimating equation for $\eta = (\beta, \gamma)'$:

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

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

The information matrix, $\mathcal{I}$, may also be defined as in the standard marginal rates

model, such that:

$$\begin{aligned}\mathcal{I}(\eta) &= -\frac{\partial^2 \ell(\eta,t)}{\partial\eta\partial\eta'} \\ &= \sum_{i=1}^{n}\int_0^\tau \left[\frac{\sum_{j=1}^n Y_j(u)\tilde{\mathbf{Z}}_j^{\otimes 2}(u)\exp\{\eta'\tilde{\mathbf{Z}}_j(u)\}}{\sum_{j=1}^n Y_j(u)\exp\{\eta'\tilde{\mathbf{Z}}_j(u)\}}\right. \\ &= \left. -\frac{\left(\sum_{j=1}^n Y_j(u)\tilde{\mathbf{Z}}_j(u)\exp\{\eta'\tilde{\mathbf{Z}}_j(u)\}\right)^{\otimes 2}}{\left(\sum_{j=1}^n Y_j(u)\exp\{\eta'\tilde{\mathbf{Z}}_j(u)\}\right)^2}\right] dN_i(u)\end{aligned}$$

Considering the developments presented in Lin et al(2000), $\hat{\mathcal{I}}(\eta)^{-1}$ and $\hat{\Gamma} = \hat{\mathcal{I}}^{-1}\hat{\Sigma}\hat{\mathcal{I}}^{-1}$ are referred, respectively, as the naive and robust covariance matrix estimators, where

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^\tau \left\{\tilde{\mathbf{Z}}_i(u) - \frac{\mathbf{S}^{(1)}(\eta,u)}{S^{(0)}(\eta,u)}\right\} d\hat{M}_i(u)\right]^{\otimes 2},$$

for $0 \leq t \leq \tau$ and $d\hat{M}_i(t) = dN_i(t) - \int_0^t Y_i(s)\exp\{\eta'\tilde{\mathbf{Z}}_i(s)\}d\hat{\mu}_0(s)$. The baseline mean is estimated by the Breslow-type estimator as $\hat{\mu}_0(t) = n^{-1}\int_0^t dN_\cdot(u)/S^{(0)}(\beta,u)$, where $dN_\cdot(u) = \sum_{i=1}^n dN_i(u)$.

Therefore, the nonparametric function $\theta(t)$ can be estimated by $\hat{\theta}(t) = \hat{\gamma}_0 + \sum_{k=1}^{m+3}\hat{\gamma}_k\tilde{B}_k(t)$ and its variance estimator is given by $var\{\hat{\theta}(t)\} = \tilde{\mathbf{B}}^*(t)'cov(\hat{\gamma})\tilde{\mathbf{B}}^*(t)$, where $\tilde{\mathbf{B}}^*(t) = (1,\tilde{B}_1(t),...,\tilde{B}_{m+3}(t))'$ 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 hypotheses tests proposed here are based on large-sample theory of maximum partial likelihood estimation (Andersen and Gill, 1982) and modern empirical process theory considered by Lin et al (2000). The theory holds if the B-spline basis is chosen a priori. In this case, the definition of the basis for the fixed model is only dependent on the data to the extent that m interior knots are placed at the quantiles of the sample distribution of the event times (Sleeper and Harrington, 1990, Abrahamowicz et al, 1996, Giorgi et al, 2003).

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 $\tilde{\gamma} =$

$(\gamma_1, ..., \gamma_{m+3})'$. Thus, the hypothesis of interest is that H_o: $\tilde{\gamma} = \mathbf{0}$. By analogy with the usual parametric likelihood procedures, this statistic can be defined by:

$$Q_W = \hat{\tilde{\gamma}}' cov(\hat{\tilde{\gamma}})^{-1} \hat{\tilde{\gamma}}$$

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.

3.2.1. Selection of Number and Location of Knots

This proposed method may be, however, sensitive to the number and location of knots. As in other methods that incorporate regression splines (Rosenberg, 1995; Abrahamowicz et al, 1996; Giorgi et al, 2003), we consider primarily splines with a fixed small number of knots. The location of the interior knots is usually based on the quantiles of the observed event times in order to ensure an approximate equal number of events in each interval (Giorgi et al, 2003). According to Abrahamowicz et al (1996), this should have a minor impact on the accuracy of inference.

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 (Sleeper and Harrington, 1990; Abrahamovicz et al, 1996). However, in such cases, additional variance is expected due to the posteriori model selection, which may inflate type I error rates. Among the criteria for choosing the proper number of knots that were proposed for univariate time-to-event settings are the generalized cross-validation (GCV) and Akaike's information criterion (AIC) (O'Sullivan, 1988, Rosenberg, 1995, Nan et al, 2003). Considering AIC, we specify several values of interior knots (m) and choose the m that minimizes $AIC(m) = -2l(\beta, \gamma) + 2(m + degree + 1)$, with degree=3 for cubic splines. The GCV method may be computationally onerous

for moderate sample sizes. However, we also discuss here an adaptation of the GCV criterion, that was proposed for univariate time-to-event data by Nan et al (2003), to the recurrent event setting. This approach also extends the method proposed by O'Sullivan (1988), in which the baseline cumulative hazard function was considered known.

The GCV method is defined as follows. Using the time-varying coefficient rates models defined previously, calculate the cumulative rate function $\hat{\mu}_0(t;m)$ and $\hat{\theta}(t;m)$ for a range of interior knots m. If $(\hat{\beta}_l, \hat{\gamma}_l)$ are the estimators of (β, γ) at the lth iteration, then the working dependent variable y_{ij} and the working weight v_{ij} for the jth observation of the ith subject can be written, respectively, as

$$y_{ij} = \hat{\tilde{\gamma}}_l' \tilde{\mathbf{W}}_i(t_{ij}) + \hat{\beta}_l' \mathbf{Z}_i(t_{ij}) + dN_i(t_{ij})/(2v_{ij}) - 1$$

and

$$v_{ij} = \frac{1}{2}\mu_0(t_{ij}) \exp\{\hat{\tilde{\gamma}}_l' \tilde{\mathbf{W}}_i(t_{ij}) + \hat{\beta}_l' \mathbf{Z}_i(t_{ij})\}.$$

Let $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_n)'$, $\hat{\mathbf{V}} = diag(\hat{\mathbf{v}}_1, ..., \hat{\mathbf{v}}_n)$ and $\hat{\mathbf{f}} = (\hat{\mathbf{f}}_1, ..., \hat{\mathbf{f}}_n)'$ be, respectively, the working dependent variable, the working weight matrix and the predicted value vector at convergence. Then $\hat{\mathbf{f}}$ can be calculated as $\hat{\mathbf{f}} = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\hat{\mathbf{V}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\hat{\mathbf{V}}\hat{\mathbf{y}} = \hat{\mathbf{H}}\hat{\mathbf{y}}$.

Considering the definitions above, calculate and plug them into the GCV equation to select the m that minimizes GCV(m), say m^*. The GCV is defined as:

$$GCV(m) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} v_{ij}(\hat{y}_{ij} - \hat{f}_{ij})^2}{(1-\bar{h})^2},$$

where $\bar{h}$ is the average of the diagonal elements of $\hat{\mathbf{H}}$ and n_i denotes the number of observations for subject i.

Considering the selected m from the previous step, replace $\mu_0(t)$ by this estimated $\hat{\mu}_0(t, m^*)$ and then treat it as fixed and known. Note that a common $\hat{\mu}_0(t, m^*)$ is used to calculate GCV for different choices of m. Then recalculate GCV using an iterative weighted least square algorithm for each choice of m and select a new m

that minimizes GCV(m). For this procedure $(\hat{\beta}_{l+1}, \hat{\gamma}_{l+1})$ is calculated by minimizing $\sum_i^n \sum_j^{n_i} v_{ij}\{y_{ij} - \tilde{\gamma}'\tilde{\mathbf{W}}_i(t_{ij}) - \beta'\mathbf{Z}_i(t_{ij})\}^2$. This last part of the process is repeated until the chosen m^* at the current step is the same as the m^* at the previous step.

3.3 Simulation Studies

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 marginal rates model:

$$E[dN(t)|\mathbf{Z}] = d\mu(t) = \exp\{\theta(t)'\mathbf{Z}\}d\mu_0(t),$$

We generated recurrent event times using the following random-effect intensity model $E[dN(t)|N(t-), \mathbf{Z}, u] = \lambda(t|Z, u) = u\lambda_0(t)\exp(\theta(t)\mathbf{Z})$, 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 $\sigma^2 = 0$ and 1. Since u is independent of Z, and E(u)=1 then the random-effect intensity model implies the marginal rates model with $d\mu_0(t) = \lambda_0(t)dt$.

We considered a constant baseline hazard function λ_0 for all configurations described here. The subject's follow-up time was uniform[0,1] and the value for λ_0 varied for the different configurations considered, such that an average of approximately 3.5 events were observed per subject during the trial period. The failure indicator Δ_{ij} was defined as $\Delta_{ij} = I(T_{ij} \leq C_i)$.

We considered three different functions for the true log of the rate ratio of the two groups as functions of time, specifically $\theta(t) = -1.2$ (a constant rate ratio over time), $log(t+1)$ and $1.2sin(-\pi t)$. We refer to these models as constant, increasing and rise/fall.

The recurrent event times were generated considering the relationships between

between $\lambda(t|Z,u)$, $\Lambda(t|Z,u)$ and $S(t|Z,u)$, denoting respectively the intensity function, the cumulative intensive function and the survival function, such that:

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

and

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

The time to the jth recurrent event was then be defined by $S_{T_j|T_{j-1},T_{j-2},\ldots T_1}(t|Z,u)$

$$\begin{aligned} &= \exp\left\{-\left[\int_0^t u\lambda_0 \exp\{\theta(t)Z\}ds - \int_0^{T_{j-1}} \lambda_0 \exp\{\theta(t)Z\}ds\right]\right\} \\ &= \exp\{-\int_{T_{j-1}}^t u\lambda_0 \exp\{\theta(t)Z\}ds\} = \zeta \end{aligned}$$

where $\zeta \sim$ Unif(0,1), j=1,2,...J_i, $T_0 = 0$ and $t > T_{j-1}$.

We have explicit solutions for the recurrent times when $\theta(t) = -1.2$ and $\theta(t) = log(t+1)$. In such cases, the time to the jth recurrent event was defined, respectively, as

$$T_j = T_{j-1} - \frac{\log \zeta}{\lambda_0 u \exp\{-1.2Z\}}$$

and

$$T_j = \sqrt[Z+1]{(T_{j-1}+1)^{Z+1} - \frac{(Z+1)\log\zeta}{\lambda_0 u}} - 1.$$

In these two cases, we considered a constant baseline hazard function $\lambda_0(t)$ equal to 10 and 6, respectively. The expression for computing the recurrent event times for $\theta(t) = 1.2sin(-\pi t)$, however, does not have a closed form. Thus, in this case we used a Newton-Raphson algorithm to obtain the recurrent event times, with $\lambda_0(t) = 10$.

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

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

The second model specifies

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

where $\tilde{A}_1 ... \tilde{A}_{m+2}$ 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 $\tilde{C}_1 ... \tilde{C}_{m+2}$ 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 (Gray, 1992; Abrahamowicz et al, 1996).

For each combination defined by the model complexity and shape of the true rate function, 1,000 samples of size 100 were generated. For each configuration, we present the sampling bias, estimated standard error (ESE), mean of the standard error of the estimates(SEE) of $\theta(t)$ and the coverage probability (CP) of the Wald 95% confidence interval. The sampling bias and sampling variance of the estimates of $\theta(t)$ are defined, respectively, as the average bias and the variance from the 1,000 random samples. Let $\hat{\theta}_i(t)$ be the estimate of the ith random sample at time t, then:

Sampling bias (t)= $\frac{\sum_{i=1}^{1000} \hat{\theta}_i(t)}{1{,}000} - \theta(t)$, Sampling variance (t) =$\frac{\sum_{i=1}^{1000} (\hat{\theta}_i(t) - \bar{\theta}(t))^2}{1{,}000}$,
where $\bar{\theta}(t) = \frac{1}{1{,}000} \sum_{i=1}^{1000} \hat{\theta}_i(t)$.

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. For estimating power 1,000 samples were used for different spline models with 2 interior knots. The simulation studies were implemented using R version 1.9.1 software.

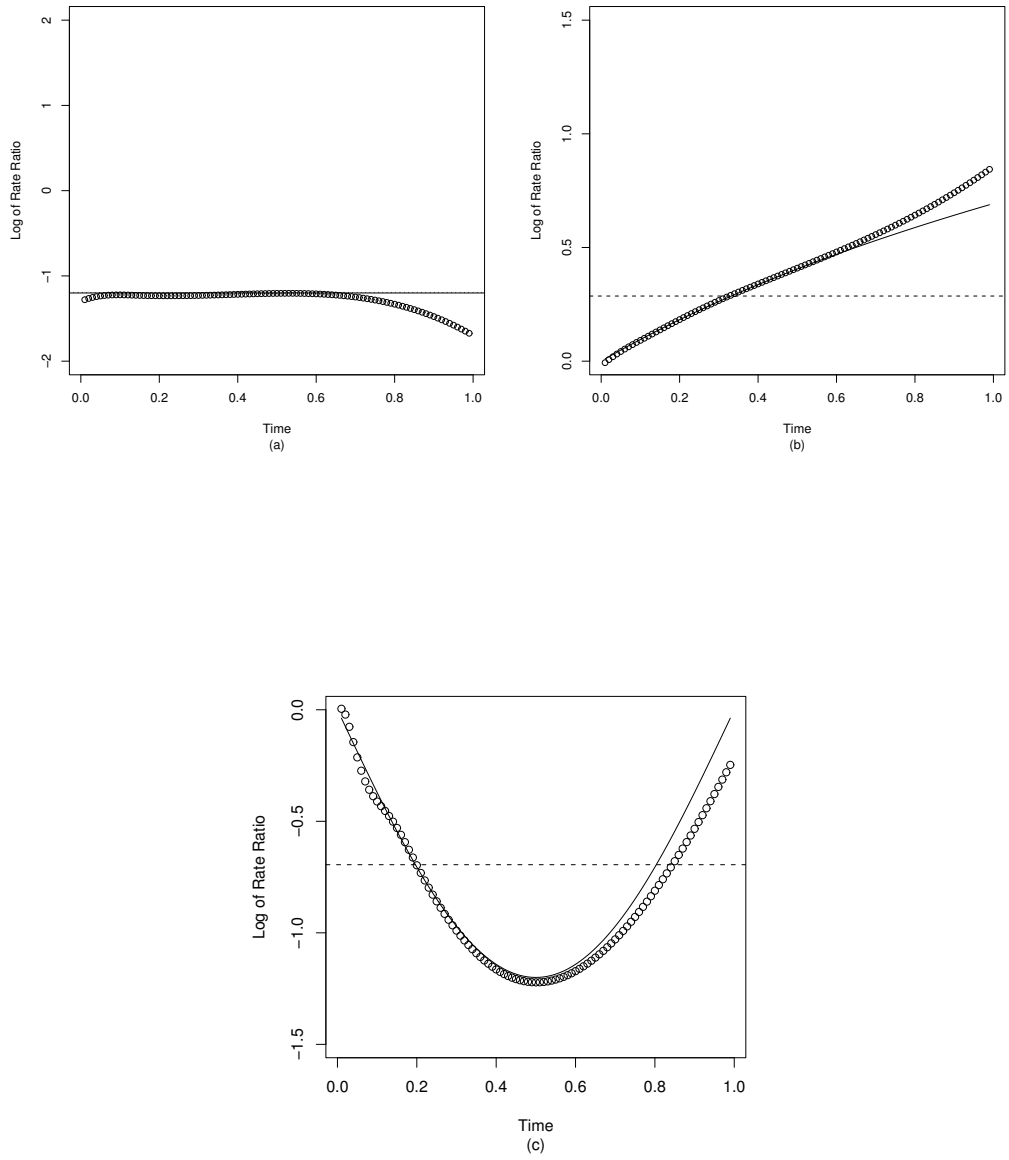

FIGURE 3.1: True Log Rate Ratio Functions (Solid Lines) against the Mean of 1,000 Estimates Using Proposed Model (Dotted Lines) for Models using different functional forms of $\theta(t)$: (a) $\theta = -1.2$, (b) $\theta = log(t+1)$, (c) $\theta(t) = 1.2sin(-\pi t)$. In each of the three panels, dashed lines represent the mean of 1,000 estimates for the standard marginal rates model.

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 in a model when it is present. Figure 3.1 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 $\sigma^2 = 0$. Similar results, that are not shown here, were obtained when $\sigma^2 = 1$. 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 will poorly describe the effects for the increasing and rise/fall models. Note in Figure 3.1 (a) 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 at the end of the study period, which is probably due to the very few events observed during this period.

TABLE 3.1: Bias, ESE, SEE and CP for estimated time-varying coefficient at selected time t for $\theta(t) = -1.2$ and n=100. The regression splines have 2 knots and the location of knots are based on the quantiles of the distribution of the recurrence times.

Spline Model	Selected Time (t)	Bias	ESE	SEE	CP
Piecewise/Linear Spline					
	0.25	0.0162	0.1990	0.2106	93.8%
	0.50	0.0172	0.1782	0.1916	93.2%
	0.75	0.0348	0.3070	0.3289	93.3%
Quadratic Spline					
	0.25	0.0281	0.2638	0.2724	94.0%
	0.50	0.0061	0.2419	0.2585	92.8%
	0.75	0.0614	0.3181	0.3681	91.7%
Cubic Spline					
	0.25	0.0327	0.2403	0.2435	94.5%
	0.50	0.0050	0.2385	0.2573	92.9%
	0.75	0.0831	0.3699	0.4425	92.2%

The simulation results for $\theta(t) = -1.2$ are presented in Table 3.1. Results from the linear spline, quadratic spline and cubic spline models are included. The bias is small, especially for t=0.25 and t=0.50. One potential reason for the increased bias for t=0.75 is that there are very few observed events later in time, which is also reflected on the larger ESE. The piecewise/linear spline model has the overall best performance for this configuration, which assumes a constant rate ratio over time. It performs well once the standard error estimates (SEE) and the empirical standard error of the estimates of $\theta(t)$ are similar and the coverage probabilities are close to 95%.

TABLE 3.2: Bias, ESE, SEE and CP for estimated time-varying coefficient in rates models with regression splines considering 2 knots with $\theta(t) = log(1+t)$.

Sample Size	Spline Model	Selected Time(t)	Bias	ESE	SEE	CP
n=100	Piecewise/Linear Spline	0.25	-0.0020	0.1597	0.1640	93.2%
		0.50	-0.0036	0.1477	0.1516	94.4%
		0.75	-0.0277	0.2416	0.2563	93.9%
	Quadratic Spline	0.25	-0.0024	0.2125	0.2142	94.5%
		0.50	-0.0042	0.1936	0.2020	93.9%
		0.75	-0.0336	0.2414	0.2735	92.0%
	Cubic Spline	0.25	-0.0031	0.2063	0.2099	95.0%
		0.50	-0.0038	0.1946	0.2000	94.0%
		0.75	-0.0386	0.2879	0.3160	92.8%
n=200	Piecewise/Linear Spline	0.25	-0.0001	0.1109	0.1110	95.0%
		0.50	-0.0046	0.1039	0.1072	95.1%
		0.75	-0.0006	0.1685	0.1773	93.1%
	Quadratic Spline	0.25	-0.0041	0.1514	0.1493	95.3%
		0.50	0.0059	0.1359	0.1370	94.2%
		0.75	-0.0049	0.1668	0.1795	92.7%
	Cubic Spline	0.25	0.0007	0.1462	0.1441	95.7%
		0.50	0.0021	0.1371	0.1378	94.3%
		0.75	0.0015	0.2003	0.2057	93.9%

In Table 3.2, 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 3.3 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.

TABLE 3.3: AIC for rates models with time-varying coefficient using regression splines with $\theta(t) = log(t+1)$ for different curves and number of knots (n=100). Location of knots are based on the quantiles of the distribution.

Spline	Number of knots				
Model	m=2	m=3	m=4	m=5	m=6
Piecewise/Linear	2847	2844	2849	2850	2852
Quadratic	2848	2849	2850	2851	2851
Cubic	2849	2850	2851	2851	2851

For the third configuration we considered the aforementioned B-splines models for $\theta(t) = 1.2sin(-\pi t)$. The results for these simulation studies are displayed in Table 3.4. The estimator of $\theta(t)$ presents small bias, particularly under the 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 3.4: Bias, ESE, SEE and CP for estimated time-varying coefficient in rates models with regression splines considering 2 knots with $\theta(t) = 1.2sin(-\pi t)$ and n=100. Location of knots are based on the quantiles of the distribution.

Spline Model	Selected Time(t)	Bias	ESE	SEE	CP
Piecewise/Linear Spline					
	0.25	0.1533	0.2072	0.2109	89.2%
	0.50	-0.1716	0.1594	0.1572	80.5%
	0.75	0.0804	0.2744	0.3066	93.1%
Quadratic Spline					
	0.25	0.0043	0.2152	0.2215	94.6%
	0.50	0.0185	0.2296	0.2297	94.7%
	0.75	0.0621	0.2653	0.3093	92.2%
Cubic Spline					
	0.25	0.0136	0.1957	0.2010	95.2%
	0.50	0.0202	0.2243	0.2230	94.6%
	0.75	0.0555	0.3181	0.3181	92.2%

Tables 3.5 and 3.6 display simulated empirical sizes and powers 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 3.5 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 it 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. Table 3.6 gives powers for tests of linearity, $\tilde{\gamma} = \mathbf{0}$, for different alternatives and spline models. The power is pretty small with smaller sample size of the alternative $log(t+1)$. Power was higher for the other alternative $1.2sin(-\pi t)$, especially for larger samples sizes.

TABLE 3.5: Empirical sizes(%) of nominal 5% Wald tests for time-dependent effects in the rates models with regression splines. Location of knots are based on the quantiles of the distribution.

Number	Sample	Cubic spline		Quadratic spline		Linear spline	
of knots	size	df	size	df	size	df	size
2	100	5	11.20	4	9.80	3	7.40
	200	5	6.60	4	6.55	3	5.55
	300	5	6.50	4	6.10	3	4.75
5	100	8	15.20	7	12.10	6	9.45
	200	8	8.20	7	6.80	6	5.75
	300	8	6.60	7	6.70	6	6.15

TABLE 3.6: Estimated powers for Wald tests for time-dependent effects in the rates models with regression splines. Location of knots are based on the quantiles of the distribution.

Alternative	Sample size	Cubic spline	Quadratic spline	Linear spline
$log(t+1)$	100	0.277	0.277	0.284
	200	0.399	0.427	0.464
	300	0.571	0.610	0.640
$1.2sin(-\pi t)$	50	0.623	0.592	0.548
	100	0.845	0.863	0.843

3.4 Application

3.4.1. Results related to the first treatment cycle

Considering the data from the vitamin A community trial described in Chapter 2, we initially applied the proposed method for the analysis of the effect of vitamin A supplementation on the occurrence of diarrheal episodes during the first treatment cycle. For all analysis presented here and in the next Section we are considering a cubic B-spline model. In order to define the appropriate number of knots to be considered in this analysis, we computed AIC and GCV criteria for a range of number of knots (m=2,...,6) (Table 3.7). According to these results, the best model for the data related to the first treatment cycle would consider 2 interior knots, regardless of the criteria.

Figure 3.2 displays the behavior of the effect of vitamin A supplementation during the first treatment cycle considering a cubic B-spline model with 2 interior knots. The rate of diarrheal episodes in the supplemented group begins to fall down after few days

of the first treatment when compared to the rate of diarrhea in the placebo group. The bigger differences in those rates is observed after 30 days of the first supplementation and prolongs until about 2 and 1/2 months after it, when the rate ratio of diarrhea episodes among the two groups starts to increase and gets close to 1 again after the third month.

TABLE 3.7: AIC and GCV for rates models with regression cubic splines considering different number of knots during first dosage cycle. Location of knots based on the quantiles of the distribution

Number of knots	AIC	GCV
2	42,602.26	84,140.83
3	42,604.33	84,393.80
4	42,605.78	84,714.50
5	42,605.45	84,629.65
6	42,607.77	84,389.15

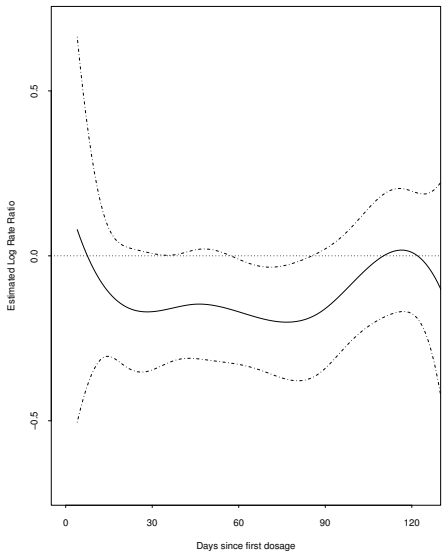

FIGURE 3.2: Estimated Log Rate Ratio Functions (solid curves) and corresponding 95% confidence intervals (dashed/dotted lines) for the effect of Vitamin A during the first treatment cycle.

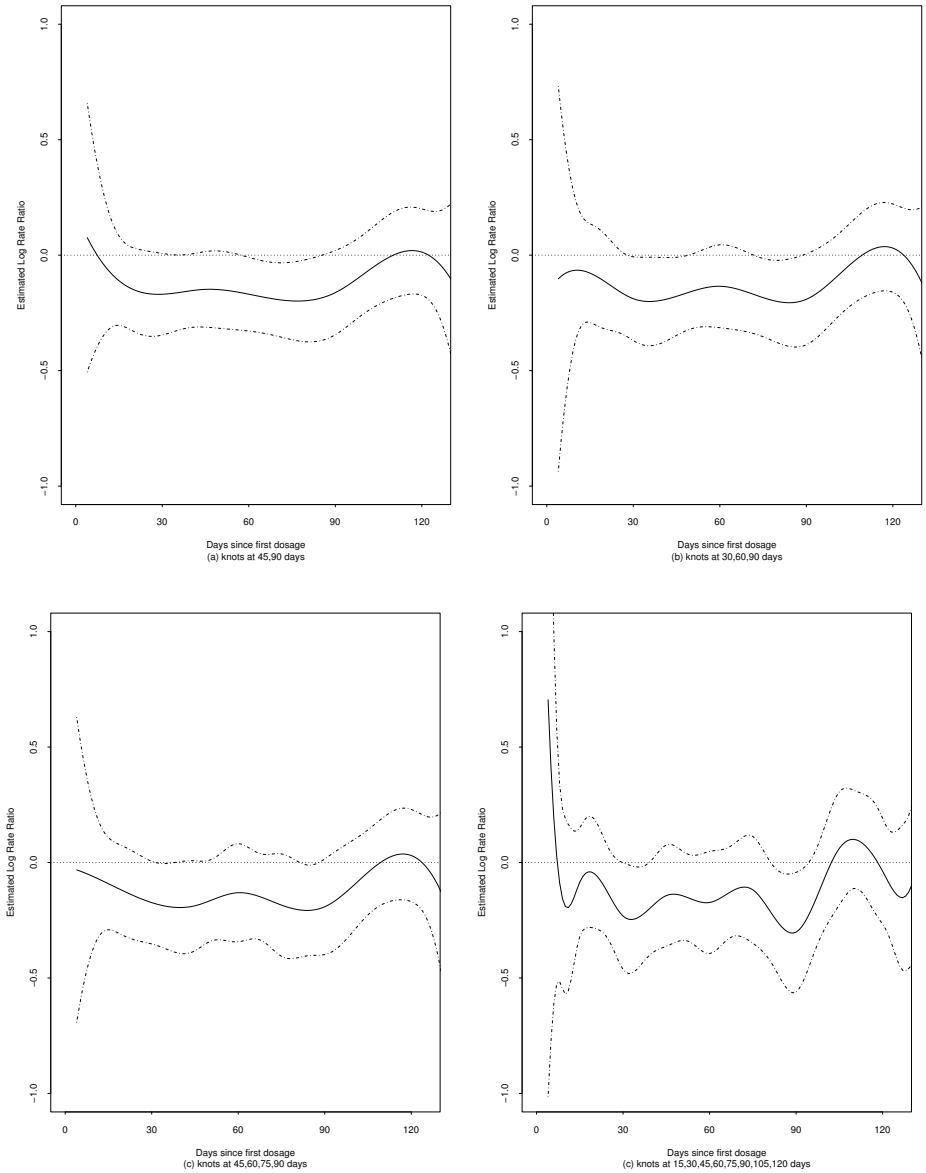

FIGURE 3.3: Estimated Log Rate Ratio Functions (solid curves) and corresponding 95% confidence intervals (dashed/dotted lines) for the effect of Vitamin A considering different knots locations.

We also estimated the treatment effect by considering pre-specified knots locations and by varying the number of knots. The cubic B-spline estimates for four different configurations are presented in Figure 3.3. The number of knots seems to impact more on the estimation of the effect of vitamin A supplementation on diarrhea over time than the knots location. The first three Figures (3.3 (a),(b) and (c)) do not seem to differ substantially. On the other hand, increasing the number of knots to 8 in Figure 3.3 (d) seems to under-smooth the relationship.

3.4.2. Results related to all treatment cycles

For the evaluation of the effect of vitamin A supplementation on diarrhea considering the information available for the entire period of the study, we first applied the standard marginal rates model, without time-dependent effects, to this data (Lin et al, 2000). The outcome of interest is the time since receiving first dose of vitamin A until the occurrence of an episode of diarrhea. 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 for gender and age at baseline. However, this overall effect was only borderline statistically significant ($\hat{\beta}$= -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 children become older (Table 3.8).

TABLE 3.8: Estimates for evaluation of treatment effect in the Vitamin A trial after adjusting for gender and age at baseline

Effects	$\hat{\beta}$	Estimated robust SE($\hat{\beta}$)	$\{\hat{\beta}/SE(\hat{\beta})\}^2$
Treatment	-0.0922	0.0501	3.382
Age	-0.0329	0.0021	248.423
Gender	0.0294	0.0503	0.340

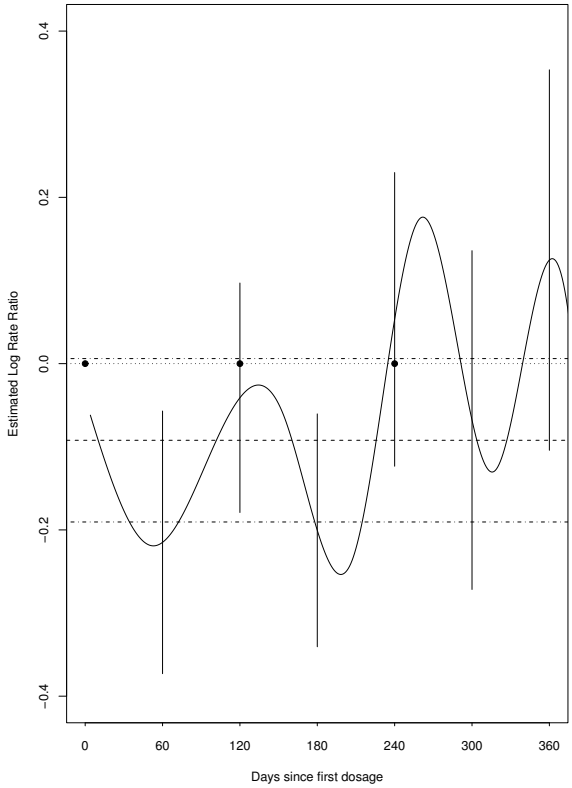

FIGURE 3.4: Estimated Log Rate Ratio Functions (solid curves) and corresponding pointwise 95% confidence intervals (vertical bars) at selected times for the Vitamin A Trial. Each panel compares the B-spline estimate with the estimate from standard marginal rate model (dashed lines) and its 95% confidence intervals (dashed-dotted lines).Dots indicate supplementation times.

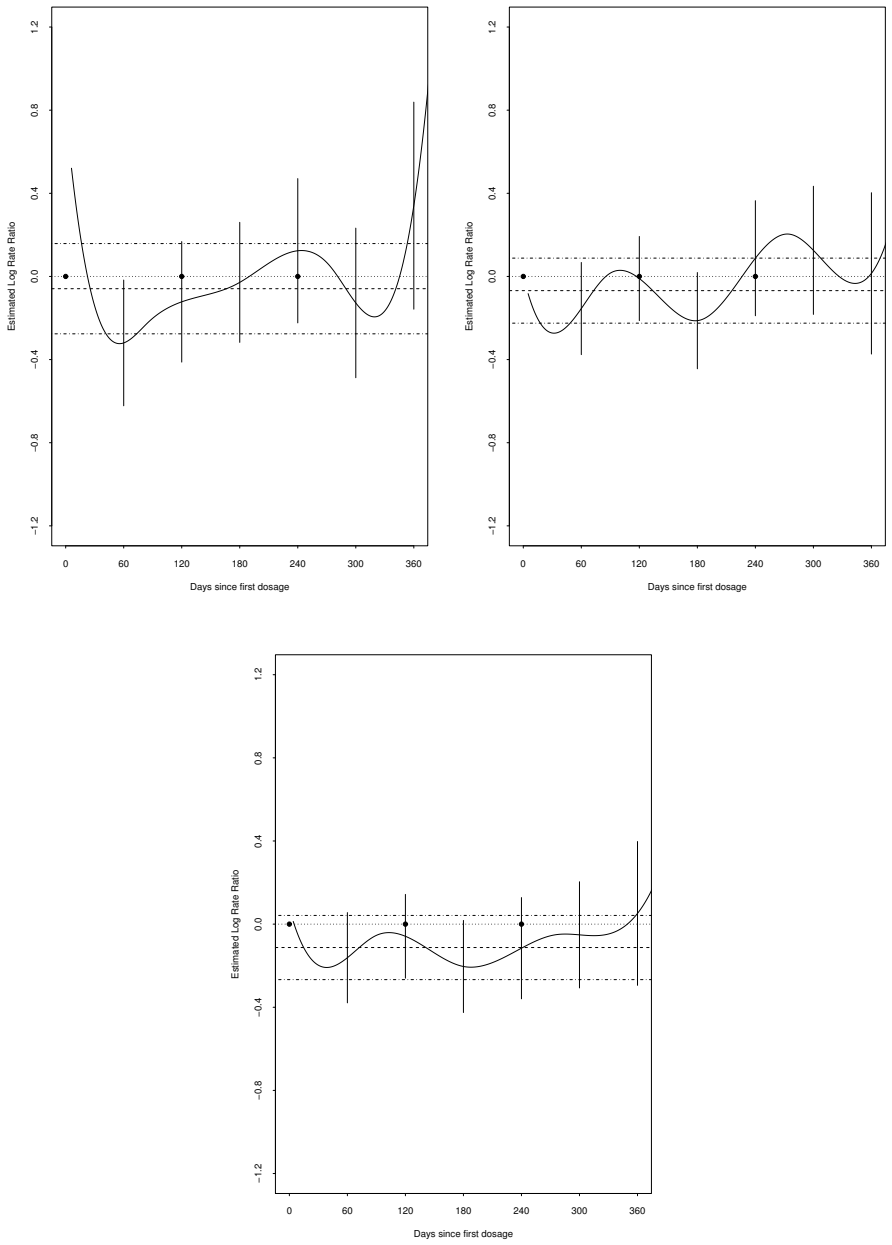

FIGURE 3.5: Estimated Log Rate Ratio Functions (solid curves) and corresponding pointwise 95% confidence intervals (vertical bars) at selected times for three age groups at baseline : (a) children with age less or equal than 12 months, (b) children with age between 12 and 24 months, (c) children older than 24 months. Each panel compare the B-spline estimate with estimate from standard marginal rate model (dotted lines) and its 95% confidence intervals (dashed lines). Dots indicate supplementation times.

In order to evaluate and describe how the effect of vitamin A supplementation behaves over time, we implemented the proposed cubic B-spline model with a larger number of knots than used previously to take into account the repeated dosages of vitamin A that the children received during the study. Figure 3.4 contains the curve for the log of the rate ratio of the occurrence of diarrhea smoothed over time considering 6 interior knots for all children in the study. 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. Results from Wald test suggest a significant time-dependent effect of treatment on the occurrence of diarrhea (p=0.0118).

TABLE 3.9: Average number of recurrent diarrheal episodes and its standard deviation by age groups. Number of children in each age group is presented in parenthesis.

Age	Vitamin A		Placebo		Overall	
group	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
≤ 12 ($n = 153$)	8.7	5.9	9.4	6.2	9.0	6.1
$12 \dashv 24$ ($n = 333$)	7.6	5.6	7.7	5.4	7.6	5.5
> 24 ($n = 721$)	4.2	4.5	4.6	4.8	4.4	4.7

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. Table 3.9 displays the average number of diarrheal episodes for each of the age groups. We verified a decreasing trend on the overall average number of recurrences, which went from 9.0 for the youngest group (age ≤ 12 months at baseline) to 4.4 for the oldest group (age > 24 months at baseline). In Figure 3.5 we present the estimates of treatment effect

over time by age group. Note from Figure 3.5 that the effect of the supplementation behaves somewhat differently among the three age groups. According to Figure 3.5, 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 effect on any of the models considered.

Table 3.9 displays the values of AIC and GCV for the analysis related to the data from children younger than 12 months at baseline. Using the former criteria, we would consider 2 and 3 interior knots, respectively, in our final model. Note, however, that the estimated trajectories of treatment effect do not differ considerably when using 2 or 3 knots for the rates model with regression B-splines for children younger than 12 months (Figure 3.6). In Figure 3.6, we present the estimates for 2, 3 and 6 interior knots with cubic B-splines. Even for the model with the larger number of knots (m=6), the behavior of the treatment effect over time do not vary drastically from the other models.

TABLE 3.10: AIC and GCV for rates models with regression cubic B-splines considering different number of knots for data of children younger than 12 months at baseline. Location of knots based on the quantiles of the distribution

Number of knots	AIC	GCV
2	13,207.05	69,282.84
3	13,207.69	69,246.40
4	13,208.15	69,270.90
5	13,207.71	69,308.69
6	13,208.82	69,268.51

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 since

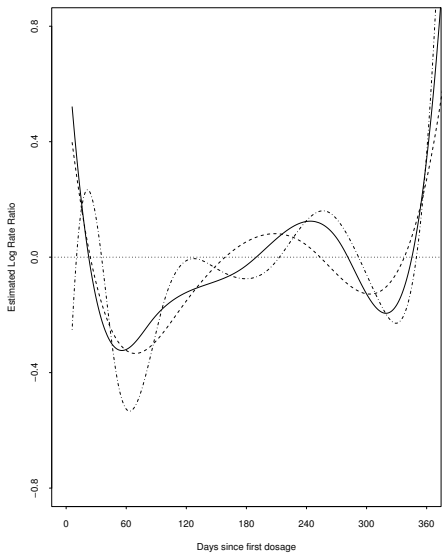

FIGURE 3.6: Estimated Log Rate Ratio Functions for children younger than 12 months using 2 knots (dashed lines), 3 knots(solid curves) and 6 knots (dotted/dashed lines) .

first dosage 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 effect again. The rate ratio for the occurrence of severe episodes of diarrhea for children 12 months when compared to children 48 months at baseline was 5.4. Figure 3.7 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. Location of knots are based on the quantiles of the distribution of the recurrence times. The reduction on the rate of severe episodes of diarrhea seems to happen earlier than that observed for any episode (Figure 3.4) after the supplementation of the first dosage of vitamin A. The treatment effect also seems subject to more variability for severe episodes than that observed for all episodes, which could be 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 events and children with recurrent severe episodes of diarrhea.

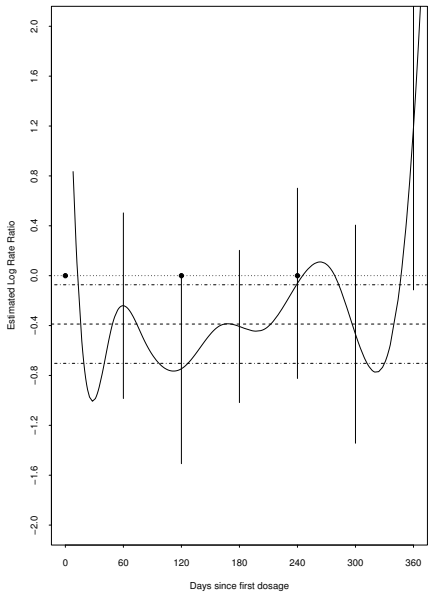

FIGURE 3.7: Estimated Log Rate Ratio Functions for severe episodes(solid curves) and corresponding pointwise 95% confidence intervals (vertical bars) at selected times for the Vitamin A Trial. Each panel compares the B-spline estimate for severe episodes with the estimate from standard marginal rate model (dashed lines) and its 95% confidence intervals (dashed-dotted lines).Dots indicate supplementation times.

3.5 Discussion

Several investigators (Sleeper and Harrington, 1990; Gray, 1992; Hastie and Tibshirani, 1993; Abrahamowicz et al, 1996; Giorgi et al, 2003; Nan et al, 2003) 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 (Giorgi et al, 2003). All available methods in the literature, however, were defined for univariate time-to-event setting. The proposed

approach is useful in estimating time-varying coefficients in the recurrent time-to-event data settings.

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 develop a method that allows the investigator to describe with details the behavior of the effects of interest over time to the rate of event occurrence. The results of simulations indicate that unless there are very few events, our 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 functions available in R. As in all survival models that require the introduction of time-varying effects, a drawback of the proposed approach is the general syntax for handling time-varying covariates. To accomodate time-varying effects, values of the covariates need to change when there is an event and thus, time intervals used for the counting process notation can break up as finely as necessary. However, the method does not require much computer time and the model can be fitted easily 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 on the 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 (Sleeper and Harrington, 1990). However, as pointed out by many authors (Sleeper and Harrigton, 1990; Gray, 1992; Giorgi et al, 2003), 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 (Giorgi et al, 2003). Some criteria for model selection has been proposed in the context of standard Cox regression, which includes cross-validation (CV) and generalized cross-validation

(GCV) criteria (O'Sullivan, 1988; Nan et al, 2003). In this paper, we adapted the GCV criterion proposed by those authors for the context of the marginal rates models with time-varying coefficients. In large data sets, however, choosing the number of knots using the GCV method can require considerable computational resources and may even be not feasible in some situations. Other criterion that has been used for the selection of number of knots considering Cox models is the Akaike information criterion (AIC). According to Abrahamowicz et al (1996), AIC-based model selection may improve the accuracy of the estimates. In this study we used both criteria to select the number of knots for the models when needed. For the selection of knots locations, it is usually appropriate to put roughly equal number of events between each of the knots (Gray, 1992; Abrahamowicz et al, 1996; Valenta and Weissfeld, 2002; Giorgi et al, 2003).

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 (Barreto et al, 1994). 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 has been showed. Studies conducted in India (Biswas et al, 1994; Chowdhury et al, 2002); China (Lie et al., 1993), Bangladesh (Rahman et al., 2001) and Brazil (Barreto et al, 1994) showed some evidence of significant reduction in overall incidence of diarrhea. At the same time, other studies (Abjeljaber et al, 1991; Ramakrishnan et al, 1995; Bhandari et al, 1994; Dibley et al., 1996; Ross et al, 1995) 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 (Bhandari et al., 1994). In another study, there was a

significant difference in the average duration of diarrhea per episode between the two groups (Biswas et al., 1994).

For the analysis of the effect of vitamin A supplementation on diarrhea, the implementation of the proposed method provided further evidence on the effectiveness of such policy to prevent diarrhea in young children, and provided more detailed insights into the behavior of such effect over time. Using a standard marginal rates model, we verified a borderline effect of vitamin A supplementation, with the occurrence rate of episodes of diarrhea since first dosage being 8.8% lower for the supplemented children compared to the placebo group. When considering severe episodes of diarrhea, the results pointed out for a reduction of 31.8% on the occurrence rate of such event. These results corroborate those already available in the literature. Furthermore, the use of the proposed method with regression splines for the analysis of the vitamin A study allowed the estimation of the treatment effect over time, providing curves of the rate ratio of the occurrence of diarrhea smoothed over time. These curves were very helpful in describing the detailed behavior of the supplementation of vitamin A in small children over time and in determining the potential duration of the effect for each of such dosages. 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, the proposed rates model provided a better description of the effect of supplementation of high doses of vitamin A over time on diarrhea in children by allowing the estimation of time-dependent effects through the use of regression splines. 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.

CHAPTER 4

PENALIZED SPLINES IN THE TIME-VARYING COEFFICIENT RATES MODEL

4.1 Introduction

In the last several years there has been significant research concerning inference and testing of varying coefficients in survival models (Lin and Wei, 1991, Gray, 1992, Gray, 1994, Abrahamowicz et al, 1996, Berhane and Wei, 2003, Nan et al, 2003, Therneau et al, 2003). A common approach is to consider spline methods to appropriately model relationships that may have nonlinear forms. Several authors have been discussing the use of regression splines to model varying effects in the context of univariate time-to-event data (Sleeper and Harrington, 1990, Abrahamowicz et al, 2003, Giorgi et al, 2003). Alternatively, other methods were proposed to explore the functional forms of the effects using splines with moderate number of knots and parameters are estimated from penalized partial likelihoods(Gray, 1992, Gray, 1994, Berhane and Wei, 2003, Therneau et al, 2003). The penalty functions are similar as those used for nonparametric penalized likelihood analysis. None of those methods, however, have been proposed for the analysis of recurrent time-to-event data.

Research has also been conducted for the development of estimation methods of means/rates of recurrent event in recent several years (Pepe and Cai, 1993, Lawless

and Nadeau, 1995, Lin et al, 2000). In this Chapter we propose a method that uses B-splines with only small to moderate number of knots to estimate time-varying effects in the marginal rates model with estimation based on penalized partial likelihood.

The proposed method is considered to examine the functional form of the effect of vitamin A supplementation on the rate of recurrent diarrheal episodes in small children. The data used here is from a randomized community trial conducted in Brazil between 1990 and 1991, with 1,207 children aged 6-48 months at baseline. The children received multiple high doses of vitamin A, which were taken every 4 months for one year. Several studies conducted to evaluate programs of vitamin A supplementation had focused on its overall average effect (Barreto et al, 1994, Biswas et al, 1994, Rahman et al, 2001, Chowdhury et al, 2002). However, it is of interest to summarize the information in the data about the shape of the supplementation effect over time. That is, is the effect of each dose of vitamin A on the rate of diarrheal episodes the same, or does it vary with time? Another question to be examined is how long the effect of each of those dosages on the recurrences lasts.

The remainder of this Chapter is organized as follows. Details of the proposed method and test statistic are given in Section 2. Simulation methods and results are discussed in Section 3. In Section 4, results from the analysis of vitamin A data are summarized. A discussion of the issues pertinent to the proposed method and its application is given in Section 5.

4.2 Model and Methods

Let N be the number of events that occur over the interval $[0, t]$ and $\tilde{\mathbf{Z}}(.) = [\mathbf{Z}(t), \mathbf{W}(t)]'$, where $\mathbf{Z}(t)$ denote the covariate vector with constant effect and $\mathbf{W}(t)$ denote the covariate whose effect could be changing with time. Both $\mathbf{Z}(t)$ and $\mathbf{W}(t)$ could be time-independent or time-dependent covariates. Let C denote the follow-up or censoring time and $Y(t) = I(C \geq t)$ be the at-risk indicator. Therefore, for a random sample of n subjects, the observable data consist of $\{N_i(.), Y_i(.), \tilde{\mathbf{Z}}_i(.)\}$, $i = 1, ..., n$.

Consider external time-dependent covariates (Kalbfleish and Prentice, p.280, 2002), such that:

$$E[dN(u)|\mathbf{Z}(u), \mathbf{W}(u)] = E[dN(u)|\mathbf{Z}(t), \mathbf{W}(t)], \text{ for all u, t, such that } t \geq u,$$

where $E[dN(s)|\mathbf{Z}(s), \mathbf{W}(s)]$ denotes the marginal failure rate ($d\mu(s)$). With fixed and external time-dependent varying covariates, the expected number of failures by time t can be written as:

$$\begin{aligned} E[N(t)|\mathbf{Z}(t), \mathbf{W}(t)] &= \int_0^t E[dN(u)|\mathbf{Z}(t), \mathbf{W}(t)] \\ &= \int_0^t E[dN(u)|\mathbf{Z}(u), \mathbf{W}(u)] \\ &= \mu(t) \end{aligned}$$

In such case, $\mu(t)$ models the expected number of failures in (0,t] as a function of $\mathbf{Z}(t)$ and $\mathbf{W}(t)$, facilitating the interpretation of the corresponding parameters.

Suppose now the following marginal rate model:

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

where β is a $p \times 1$ vector of fixed regression parameters, $\theta(t)$ is the time-varying regression parameter and $d\mu_0(t)$ is an unspecified baseline rate function. Thus, the corresponding mean model can be defined as

$$\mu(t) = \int_0^t \exp\{\beta' \mathbf{Z}(u) + \theta(u)\mathbf{W}(u)\} d\mu_0(u)$$

When considering internal time-dependent covariates, $\mu(t)$ may be interpreted as a cumulative rate function.

We propose to estimate the time-varying parameter $\theta(t)$ through the use of a penal-

ized B-spline based model with small to moderate number of knots (2-10). We consider, for instance, a standard cubic spline model for $\theta(t)$, such that $\theta(t) = \gamma_0 + \sum_{k=1}^{m+3} \gamma_k \tilde{B}_k(t)$, where $\tilde{B}_k(t), (k = 1, ..., m+3)$ denotes the B-spline basis functions. These functions can be defined recursively, such that the kth B-spline of order q (that is, degree q - 1) is a weighted sum of the kth and (k+1)st B-splines of order q - 1, with weights depending on the breakpoints and continuity conditions (Sleeper and Harrington, 1990).

In this case the following time-varying coefficient rate model is considered:

$$d\mu(t) = \exp\{\beta' \mathbf{Z}(t) + (\gamma_0 + \sum_{k=1}^{m+3} \gamma_k \tilde{B}_k(t))\mathbf{W}(t)\} d\mu_0(t), t \geq 0.$$

The standard form of the penalty function for cubic splines is used here, which is given by $\frac{1}{2}\alpha \int_{min(t_i)}^{max(t_i)} [\theta''(t)]^2 dt$, where α denotes the smoothing parameter. This penalty function can be rewritten as $\frac{1}{2}\alpha \tilde{\gamma}' \tilde{\mathbf{D}} \tilde{\gamma}$, where $\tilde{\gamma}' = (\gamma_1, ..., \gamma_{m+3})$ and $\tilde{\mathbf{D}}$ is an appropriately chosen symmetric, nonnegative matrix, such that the *i,j* th element of $\tilde{\mathbf{D}}$ would be $\int_{min(t_i)}^{max(t_i)} \tilde{B}_i''(t) \tilde{B}_j''(t) dt$.

The parameter estimates are obtained by maximizing the penalized log partial likelihood, which is defined as

$$\ell_p(\beta, \gamma) = \ell(\beta, \gamma) - \frac{1}{2}\alpha \gamma' \mathbf{D} \gamma,$$

where $\ell(\beta, \gamma) = \sum_{i=1}^{n} \int_0^t \{\eta' \tilde{\mathbf{Z}}_i(u) - log[S^{(0)}(\beta, \gamma, u)]\} dN_i(u)$ is the usual log partial likelihood for the marginal rates model (Lin et al, 2001), with $\gamma' = (\gamma_0, \tilde{\gamma}')$, $S^{(0)}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{\beta' \mathbf{Z_i}(t) + \gamma' \tilde{\mathbf{W}}_\mathbf{i}(t)\}$, $\tilde{\mathbf{W}}_i(t) = (\mathbf{W}_i(t), \tilde{B}_1(t)\mathbf{W}_i(t), ..., \tilde{B}_{m+3}(t)\mathbf{W}_i(t))'$, $\eta = (\beta, \gamma)'$ and $\tilde{\mathbf{Z}}_i(u) = (\mathbf{Z}_i(u), \tilde{\mathbf{W}}_i(u))'$. For the cubic splines setting, $\mathbf{D}$ is an $(m+4) \times (m+4)$ matrix with the first row and column being zeros, since the constant term passes unpenalized.

To estimate β and γ, one solves the score equations. Since the penalty function does not involve β, then the score equations for β are identical to those for the standard

marginal rates model:

$$\mathbf{U}_\beta = \sum_{i=1}^{n} \int_0^t \left[\mathbf{Z}_i(u) - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(u) \exp\{\beta' \mathbf{Z_i}(u) + \gamma' \tilde{\mathbf{W}}_\mathbf{i}(u)\} \mathbf{Z_i}(u)}{\frac{1}{n}\sum_{i=1}^{n} Y_i(u) \exp\{\beta' \mathbf{Z_i}(u) + \gamma' \tilde{\mathbf{W}}_\mathbf{i}(u)\}} \right] dN_i(u).$$

The score equations for the γ is then:

$$\mathbf{U}_\gamma = \sum_{i=1}^{n} \int_0^t \left[\tilde{\mathbf{W}}_\mathbf{i}(\mathbf{u}) - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(u) \exp\{\beta' \mathbf{Z_i}(u) + \gamma' \tilde{\mathbf{W}}_\mathbf{i}(u)\} \tilde{\mathbf{W}}_i(u)}{\frac{1}{n}\sum_{i=1}^{n} Y_i(u) \exp\{\beta' \mathbf{Z_i}(u) + \gamma' \tilde{\mathbf{W}}_\mathbf{i}(u)\}} \right] dN_i(u) - \alpha \mathbf{D} \gamma.$$

The baseline mean is estimated by the Breslow-type estimator as

$$\hat{\mu}_0(t) = n^{-1} \int_0^t dN_\cdot(u) / S^{(0)}(\beta, u),$$

where $dN_\cdot(u) = \sum_{i=1}^{n} dN_i(u)$.

The penalized partial likelihood can be fitted with the Newton-Raphson algorithm. In addition to the score vectors $\mathbf{U}_\beta$ and $\mathbf{U}_\gamma$, this requires the definition of the Hessian of the penalized partial log-likelihood:

$$\mathbf{H} = \mathcal{I} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{D} \end{pmatrix},$$

where $\mathcal{I}$ is the usual unpenalized information matrix for the marginal rates model, which is defined as

$$\begin{aligned} \mathcal{I} &= \sum_{i=1}^{n} \int_0^\tau \left[\frac{\sum_{j=1}^{n} Y_j(u) \tilde{\mathbf{Z}}_j^{\otimes 2}(u) \exp\{\eta' \tilde{\mathbf{Z}}_j(u)\}}{\sum_{j=1}^{n} Y_j(u) \exp\{\eta' \tilde{\mathbf{Z}}_j(u)\}} \right. \\ &= \left. - \frac{\left(\sum_{j=1}^{n} Y_j(u) \tilde{\mathbf{Z}}_j(u) \exp\{\eta' \tilde{\mathbf{Z}}_j(u)\}\right)^{\otimes 2}}{\left(\sum_{j=1}^{n} Y_j(u) \exp\{\eta' \tilde{\mathbf{Z}}_j(u)\}\right)^2} \right] dN_i(u). \end{aligned}$$

Following similar arguments as in Gray (1992, 1994) and Lin et al (2000), the

covariance matrix for $\hat{\eta}$ can be consistently estimated by

$$\hat{\Gamma}_p = \hat{\mathbf{V}}\hat{\Sigma}\hat{\mathbf{V}},$$

where

$\mathbf{V} = \mathbf{H}^{-1}\mathcal{I}\mathbf{H}^{-1}$ and

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^\tau \left\{\tilde{\mathbf{Z}}_i(u) - \frac{\mathbf{S}^{(1)}(\eta,u)}{S^{(0)}(\eta,u)}\right\} d\hat{M}_i(u)\right]^{\otimes 2},$$

for $0 \leq t \leq \tau$, with $\mathbf{S}^{(1)}(\eta,t) = n^{-1}\sum_{j=1}^{n} Y_j(t)\exp\{\eta'\tilde{\mathbf{Z}}_j(t)\}\tilde{\mathbf{Z}}_j(t)$ and $d\hat{M}_i(t) = dN_i(t) - \int_0^t Y_i(s)\exp\{\eta'\tilde{\mathbf{Z}}_i(s)\}d\hat{\mu}_0(s)$.

The estimator $\hat{\mathbf{\Sigma}}$ is defined in terms of the unpenalized score contributions since the penalty contributions are asymptotically negligible under the null hypothesis. Note, however, that $\hat{\Gamma_p}$ is a function of the information matrix, the penalty matrix and the smoothing parameter. The inferential procedures for the first $p+1$ elements of η are directly analogous to those outlined in Lin et al (2001) because the penalty matrix $\mathbf{D}$ contributes to the penalized score and information matrix only for the last (m + 3) components of η.

4.2.1. Test Statistic

This section considers a Wald-type statistic for hypotheses of the form $H_0 : \mathbf{C}\eta = \mathbf{0}$, where $\mathbf{C}$ has full row rank $r \leq m + 4 + Dim(\beta)$. Under H_0, the test statistic has the quadratic form

$$(C\hat{\eta})^T(C\hat{\Gamma}_p C^T)^{-1}(C\hat{\eta}) \sim \chi^2_r,$$

Constructing tests for the hypothesis that the effect is linear, i.e., H_0: $\tilde{\gamma} = \mathbf{0}$ is done in exactly the same way, with $r = m + 3$.

An alternative variance estimator to be considered in this setting is through the use of $\mathbf{H}^{-1}$ directly instead of $\mathbf{V}$ to define the covariance matrix, i.e., $\hat{\hat{\Gamma}}_p = \hat{\mathbf{H}}^{-1}\hat{\Sigma}\hat{\mathbf{H}}^{-1}$,

as discussed by Therneau et al (2003) for penalized survival models with frailty. For penalized smoothing splines in Cox regression, significance tests are based on $\mathbf{H}^{-1}$ as the most conservative choice (Therneau, Grambsch and Pankratz, 2003).

4.2.2. Choice of Smoothing Parameter and Placement of Knots

We adopted here similar approach as suggested in various papers (Gray, 1992; Buja et al, 1989, for instance) to define the smoothing parameter for our recurrent event settings. Theoretically the smoothing parameter is considered fixed and defined a priori in order to be used for obtaining the parameter estimates. However, operationally the smoothing parameter is calculated by the following relationship with the degrees of freedom, which should be specified for each nonparametric term, such that:

$$df = trace\{\lim \mathcal{I}_{\gamma\gamma|\beta}/\bar{v}(\mathcal{I}_{\gamma\gamma|\beta}/\bar{v} + \alpha\mathbf{D})^{-1}\},$$

where $\mathcal{I}_{\gamma\gamma|\beta} = \mathcal{I}_{\gamma\gamma} - \mathcal{I}_{\gamma\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\gamma}$ and $\bar{v}$ refers to the average number of recurrent events per subject.

According to Gray (1992) the number and location of knots are not very important when considering reasonably spread out knots, such that roughly equal number of data is put between the knots. This same algorithm is adopted here. Simulation studies pointed out for improved results when considering 5 degrees of freedom for several scenarios (Gray, 1992, Gray, 1994, Berhane and Weissfeld, 2003).

4.3 Simulation Studies

Simulation studies were conducted to examine the performance of the proposed procedure for conducting inference for the effect of time-varying coefficients on time to recurrent events. Here the finite-sample properties of the proposed parameter estimators and the size of the Wald-type test for a linear effect of the covariate modelled with splines were assessed. For each simulated data set, we estimated the time-varying

coefficient $\theta(t)$ under the marginal rates model:

$$d\mu(t) = \exp\{\beta' Z + \theta(t)W\} d\mu_0(t).$$

We generated recurrent event times using the following random-effect intensity model $\lambda(t|Z, W, u) = u\lambda_0(t)\exp\{\beta' Z + \theta(t)W\}$, where u is an unobserved unit-mean positive random variable that is independent of Z and W. Covariate values were generated for W from a Bernoulli distribution (0.5) and for Z from a Uniform distribution (0,1). We generated independent u_i (i=1,...,n) from gamma distribution, with mean 1 and variance $\sigma^2 = 0$ and 1. Thus, the random-effect intensity model implies the marginal rates model with $d\mu_0(t) = \lambda_0(t)dt$. We considered a constant baseline hazard function λ_0 for all configurations described here. The subject's follow-up time was uniform[0,1] and the value for λ_0 varied for the different configurations considered, such that an average of approximately 3.5 events were observed per subject during the trial period. The failure indicator Δ_{ij} was defined as $\Delta_{ij} = I(T_{ij} \leq C_i)$. The smoothing parameter was defined through its relationship with the degrees of freedom. We considered a priori fixed degrees of freedom of 5 for all configurations.

We consider three different functional forms of $\theta(t)$: (i) $\theta(t) = -1.2$, (ii) $\theta = log(t+1)$ and (iii) $\theta(t) = 1.2sin(-\pi t)$. The recurrent event times were generated analytically for $\theta(t) = -1.2$ and $\theta = log(t+1)$. In these two cases, we considered a constant baseline hazard function $\lambda_0(t)$ equal to 10 and 6, respectively. The expression for computing the recurrent event times for $\theta(t) = 1.2sin(-\pi t)$, however, does not have a closed form. Thus, in this case we used a Newton-Raphson algorithm to obtain the recurrent event times, with $\lambda_0(t) = 10$.

A cubic B-spline model was considered for $\theta(t)$, such that

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

where $\tilde{B}_1 ... \tilde{B}_{m+3}$ is a cubic B-spline basis. For comparison purposes we also considered a quadratic B-splines model

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

where $\tilde{A}_1 ... \tilde{A}_{m+2}$ is a quadratic B-spline basis, with penalty $\frac{1}{2}\alpha \int_{min(t_i)}^{max(t_i)} [\theta'(t)]^2 dt$.

The estimation for $\theta(t)$ was conducted considering different number of interior knots, sample size and shape of the true rate function. The location of the interior knots was chosen to ensure an approximately equal number of failures between the knots (Gray, 1992; Abrahamowicz et al, 1996).

For each combination defined by the model complexity and shape of the true rate function, 1,000 samples of size 100 were generated. For each configuration, we present the sampling bias, estimated standard error (ESE), mean of the standard error of the estimates(SEE) of $\theta(t)$ and the coverage probability (CP) of the Wald 95% confidence interval. The sampling bias and sampling variance of the estimates of $\theta(t)$ are defined, respectively, as the average bias and the variance from the 1,000 random samples. Let $\hat{\theta}_i(t)$ be the estimate of the ith random sample at time t, then

Sampling bias (t)= $\frac{\sum_{i=1}^{1000} \hat{\theta}_i(t)}{1{,}000} - \theta(t)$, Sampling variance (t) =$\frac{\sum_{i=1}^{1000} (\hat{\theta}_i(t) - \bar{\theta}(t))^2}{1{,}000}$,
where $\bar{\theta}(t) = \frac{1}{1{,}000} \sum_{i=1}^{1000} \hat{\theta}_i(t)$.

TABLE 4.1: Bias, ESE, SEE and CP for estimated time-varying coefficient in the rates models with penalized cubic splines at selected time t for $\theta(t) = -1.2$, considering different number of knots and sample sizes. Location of knots are based on the quantiles of the distribution.

# of	Selected	n=100				n=400			
knots	Time (t)	Bias	ESE	SEE	CP	Bias	ESE	SEE	CP
2									
	0.25	0.0157	0.2281	0.2318	93.7%	0.0118	0.1145	0.1162	94.3%
	0.50	0.0155	0.2303	0.2457	93.0%	0.0055	0.1154	0.1193	94.3%
	0.75	0.0602	0.3405	0.4084	90.4%	0.0235	0.1730	0.1851	93.3%
6									
	0.25	0.0180	0.2686	0.2987	91.8%	0.0108	0.1457	0.1619	92.7%
	0.50	0.0183	0.2141	0.2462	91.2%	0.0001	0.1115	0.1181	93.0%
	0.75	0.0409	0.3039	0.3673	90.0%	0.0171	0.1543	0.1633	94.7%
10									
	0.25	0.0186	0.2818	0.3229	90.1%	0.0015	0.1315	0.1482	91.9%
	0.50	0.0180	0.2348	0.2795	90.0%	-0.0011	0.1108	0.1250	92.0%
	0.75	0.0400	0.3031	0.3675	89.9%	0.0043	0.1480	0.1647	93.3%

Empirical sizes of the spline based tests, based on 2,000 samples, were examined under various specifications of number of knots (2,10) and samples sizes (n = 100, 200,

300 and 400) for the cubic B-spline model. The simulation studies were implemented using R version 1.9.1 software.

Under the cubic B-spline, we present results for $\theta(t) = -1.2$ with different number of knots and sample sizes (n=100 or 400) (Table 4.1). We observe bigger bias and smaller CP for the model with 10 knots. However, the results indicate a clear improvement for samples of size 400, implying smaller bias, more accurate variance estimator and more reasonable coverage probabilities of the Wald 95% confidence interval.

TABLE 4.2: Bias, ESE, SEE and CP for estimated time-varying coefficient in the rates models with penalized method at selected time t for $\theta(t) = log(t+1)$ and n=100, considering different number of knots and spline models. Location of knots are based on the quantiles of the distribution.

# of	Selected	Quadratic Splines				Cubic splines			
knots	times (t)	Bias	ESE	SEE	CP	Bias	ESE	SEE	CP
2									
	0.25	-0.0134	0.2084	0.2075	94.7%	-0.0048	0.2018	0.2037	95.5%
	0.50	-0.0020	0.1894	0.2029	92.7%	-0.0042	0.1908	0.2016	93.5%
	0.75	-0.0194	0.2372	0.2724	92.5%	-0.0139	0.2780	0.3082	92.0%
6									
	0.25	-0.0135	0.2346	0.2665	91.4%	-0.0098	0.2273	0.2496	92.9%
	0.50	-0.0082	0.1979	0.2267	90.3%	-0.0023	0.1830	0.2064	91.3%
	0.75	-0.0001	0.2254	0.2654	90.2%	-0.0139	0.2389	0.2748	91.8%
10									
	0.25	-0.0145	0.2186	0.2714	88.7%	-0.0189	0.2127	0.2427	91.4%
	0.50	-0.0094	0.1874	0.2317	88.8%	-0.0118	0.2084	0.2084	90.8%
	0.75	-0.0045	0.2069	0.2543	88.6%	-0.0047	0.2534	0.2534	91.9%

The results given in Table 4.2 indicate improved results when considering smaller number of knots (m=2) for $\theta(t) = log(t+1)$. In general, we observe small bias. The variance estimator is quite accurate as the mean of the standard error estimates (SEE) and the empirical error of the estimates of $\theta(t)$ (ESE) are quite similar. Table 4.2 also shows empirical coverage probabilities (i.e., proportion of samples in which the nominal 95% pointwise confidence interval includes the true value) for the models with $\theta(t) = log(t+1)$. For the model with 2 knots, the CPs are close to the nominal .95.

However, as the number of knots increases poorer results were observed. The cubic B-spline model outperforms the quadratic B-spline model for all configurations. In Table 4.3, the results for the quadratic and cubic B-splines models for $\theta(t) = 1.2sin(-\pi t)$ with 2 and 10 interior knots are displayed. Generally, the same patterns as in Table 4.2 are observed here. Coverage probabilities for the models with 10 knots are low.

TABLE 4.3: Bias, ESE, SEE and CP for estimated time-varying coefficient in the rates models with penalized method at selected time t for $\theta(t) = 1.2sin(-\pi t)$ and n=100, considering different number of knots and spline models. Location of knots are based on the quantiles of the distribution.

# of	Selected	Quadratic Splines				Cubic splines			
knots	Times (t)	Bias	ESE	SEE	CP	Bias	ESE	SEE	CP
2									
	0.25	0.0281	0.2215	0.2303	93.9%	0.0157	0.1890	0.1968	94.2%
	0.50	0.0048	0.2357	0.2477	94.0%	0.0048	0.2171	0.2280	93.4%
	0.75	0.0487	0.2750	0.3360	91.5%	0.0529	0.3038	0.3267	93.3%
10									
	0.25	0.0346	0.2255	0.2877	86.3%	0.0207	0.2274	0.2621	90.2%
	0.50	-0.0391	0.2069	0.2538	87.9%	-0.0246	0.1981	0.2322	89.8%
	0.75	0.0507	0.2312	0.2980	86.8%	0.0198	0.2505	0.2893	90.9%

Table 4.4 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=400).

TABLE 4.4: Empirical sizes of nominal 1%, 5% and 10% Wald tests for time-dependent effects in the rates models with penalized splines. Location of knots are based on the quantiles of the distribution.

# of knots	Sample Size	Nominal level 0.01	0.05	0.10
2	100	0.0570	0.1485	0.2275
	200	0.0275	0.0830	0.1430
	300	0.0190	0.0720	0.1355
	400	0.0180	0.0640	0.1200
6	100	0.1585	0.2895	0.3710
	200	0.0770	0.1585	0.2380
	300	0.0420	0.1140	0.1825
	400	0.0335	0.0935	0.1565
10	100	0.2315	0.3955	0.4960
	200	0.1100	0.2090	0.2865
	300	0.0505	0.1520	0.2305
	400	0.0415	0.1165	0.1715

4.4 Application

We applied the proposed methods to the aforementioned vitamin A study to evaluate the effect of multiple doses of vitamin A supplementation on the occurrence of recurrent episodes of diarrhea in small children. Data used here are from a randomized community trial, including 1,207 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. We fitted the following marginal rates model:

$$E[dN(t)|sex, age] = \exp\{\beta_1 sex + \beta_2 age + \theta(t) trt\} d\mu_0(t),$$

considering a cubic B-spline model for $\theta(t)$ as $\gamma_0 + \sum_{k=1}^{m+3} \gamma_k \tilde{B}_k(t)$, where $\tilde{B}_k(t)$,$(k = 1, ..., m+3)$ denotes the B-spline basis functions.

The main interest was the effect of vitamin A supplementation (trt), adjusted for children's age and gender. We conducted separate analysis considering as outcomes both (i) time since first supplementation of vitamin A to the occurrence of any diarrheal

episode and (ii) time since first supplementation of vitamin A to the occurrence of severe episodes of diarrhea. The estimated regression coefficients for both models are presented in Table 4.5. considering 6 interior knots and 5 degrees of freedom for the selection of the smoothing parameter. Models with 6 knots were considered based on researcher's prior knowledge of the shape of the curve and, thus allowing the estimates to capture the pattern of the effect due to multiple dosages of vitamin A. For both models, note that age is a significant factor to the occurrence of recurrent episodes of diarrhea, being negatively associated to their occurrence, while the gender of the child appears not to be an important factor.

TABLE 4.5: Estimates for evaluation of treatment effect in the Vitamin A trial after adjusting for gender and age at baseline in the penalized rates model

Outcome	Covariate	Estimate	Test Statistic	df	P-value
All Episodes					
	Age	-0.0330	248.50	1	< 0.0001
	Gender	0.0292	0.34	1	0.8643
	Treatment (linearity)		21.40	9	0.0110
Severe Episodes					
	Age	-0.0468	55.16	1	< 0.0001
	Gender	-0.2616	2.49	1	0.1100
	Treatment (linearity)		13.77	9	0.1307

Results from the Wald test, as defined in Section 4.2.1., pointed out for a significant time-dependent effect of treatment when considering any episode of diarrhea (p= 0.0110). At the same time, there is a lack of evidence of a time-varying effect of treatment on the occurrence of severe episodes (p= 0.1307).

The smooth function estimates for treatment effect considering all episodes of diarrhea are presented in Figure 4.1. Note an important reduction on the estimated log rate ratio 60 days after the supplementation of the first and second doses of vitamin A, with the benefit from each of the dosages completely disappearing about 2 months later. The effect of the third dosage of vitamin A seems not to be as large as those associated to the first two supplementations.

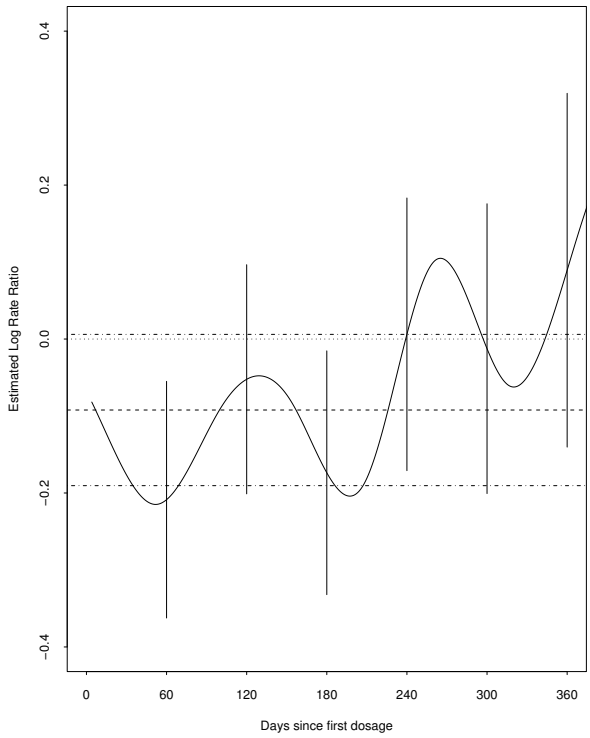

FIGURE 4.1: Estimated Log Rate Ratio Functions (solid curves) and corresponding pointwise 95% CI's at selected times (vertical lines) for the effect of Vitamin A for all episodes.

Figure 4.2 contains the smooth curve for the log of the rate ratio of the occurrence of severes episodes of diarrhea over time considering 6 interior knots. For such episodes, the effect of vitamin A supplementation behaves somewhat differently from that observed in Figure 4.1. This may imply a lasting benefit and more stable effect of vitamin A supplementation in severe episodes. A important reduction on the rate of severe episodes in the supplemented children compared to those in the placebo group was only observed after receiving the the first dose of the supplementation. This could be, however, consequence of the very low rate of incidence of severe episodes of diarrhea. In the vitamin A study the number of severe episodes of diarrhea corresponds for only 4% of all episodes during the study. The number of severe episodes of diarrhea is even smaller later in the study.

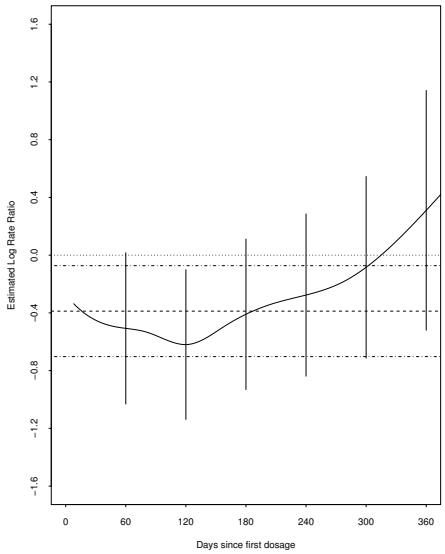

FIGURE 4.2: Estimated Log Rate Ratio Functions(solid curves) and corresponding pointwise 95% CI's at selected times(vertical lines) for the effect of Vitamin A for severe episodes.

4.5 Discussion

We propose methods for estimating time-varying coefficients in the rates models using penalized partial likelihood and cubic B-splines with small to moderate number of knots. Estimating equations are proposed for the time-varying parameter and Wald-type statistic is defined. Simulation results imply that the asymptotic properties are applicable to finite samples for the scenario with small number of knots.

The proposed methods were applied to data from a randomized community trial to evaluate the effect of vitamin A supplementation on diarrheal recurrences. Results of the analysis illustrate that the estimated rate of diarrhea occurrence reduces continuously in the supplemented group compared to placebo during the first 60 days after vitamin A supplementation. After that the benefits start to disappear. Similar patterns were observed after the first two dosages. However the effect of the third dose was not as large as those associated to the previous ones. There may be many reasons to explain such changes. First, diarrhea is a complex syndrome that might be caused by different pathogens, which includes viruses, bacteria and protozoa, and may have

different clinical forms and seasonality (Byers, Guerrant and Farr, 2001; Barreto et al, 2005). Data about pathogens associated to diarrhea episodes was not obtained in the vitamin A study, but it is expected to be equally observed in the children in the two treatment groups. It may be possible that the effect of vitamin A varies with pathogens that cause diarrhea and, therefore, changes over the year. Other important issue is that incidence of diarrhea decreases substantially with age. Thus, as expected, reduced number of episodes was observed later in the study in both placebo and vitamin A groups. Our results contribute to provide a detailed description of the effect over time of successive high doses of vitamin A, which is lacking in the literature.

In many other applications there may be prior interest in modeling and testing time-varying effect of a single predictor when adjusting for covariates whose effects are a priori known to be constant over time. Our method allows for simultaneous estimation of the effects of variables for which constant effects hold and those for which time-varying effects are expected. Our methods have the advantage of being able to estimate a relatively realistic functional form for the covariate effects of interest by using penalized B-splines, which offers an attractive compromise between fully nonparametric regression smoothers and unpenalized regression splines.

The major practical limitation of the methods proposed here is that they might be computationally intensive, particularly when increasing sample size and model complexity (number of knots and spline model). The standard form of the penalty function was considered here, which was given by the integral of a squared higher derivative of the fitted curve. Further study might propose other choices of the penalty functions that may improve small sample properties of the estimators.

CHAPTER 5

COMPARISON OF METHODS FOR DEPENDENT CENSORING: A SIMULATION STUDY

5.1 Introduction

Many approaches have been proposed to model recurrent time-to-event data, when each subject may experience repeated occurrences of the same type of event (Prentice, Williams and Petterson, 1981; Andersen and Gill, 1982; Pepe and Cai, 1993; Lin et al, 2000). Examples of recurrent data are repeated asthma attacks or recurrent pyogenic infections in chronic granulomatous disease (CGD) patients. The most commonly used methods available to model such data assume independent censoring, e.g., the censoring process is unrelated to the event failure process. However, dependent censoring arises in many studies. Ghosh and Lin (2003) suggest that the recurrent event times are subject to both independent and dependent censoring in a typical medical study and warn that the traditional methods for analysis of recurrent time-to-event data are not valid in the presence of dependent censoring. Recently, approaches to handle dependent censoring for the analysis of recurrent time-to-event data have been proposed in the literature.

Wang, Qin and Chiang (WQC) (2001) proposed to model the occurrence of recurrent events by a subject-specific nonstationary Poisson process via a latent variable, allowing the censoring mechanism be possibly informative. Their approach adopted a

multiplicative intensity function as the underlying model. WQC showed that, under regularity conditions, the resulting estimators are consistent.

Miloslavsky, Keles, van der Laan, Butler (MKLB) (2004) proposed inverse probability of censoring weighted (IPCW) estimators for the regression parameters in the Andersen-Gill model. They also extended their approach for the proportional rates model. MKLB proposed estimating equations based on IPCW mapping (Robbins and Rotnitzky, 1992) and showed that their estimators are also consistent.

Computer programs for general purpose are not available to model recurrent time-to-event data using those approaches. However, a R library for fitting the WQC model is available upon request to the WQC authors and MKLB approach can be implemented by adapting standard routines available in statistical software packages.

Besides the theoretical derivation of their estimation approaches, WQC and MKLB have also conducted simulation studies to study the finite sample properties of their estimators. WQC used 500 samples of size 400 to estimate the effect of a Bernoulli variable on the occurrence of recurrent events. They computed bias, standard errors and 95% bootstrap confidence intervals for their estimator, concluding for its validity. MKLB, on the other hand, considered 2,000 samples of size 200, with fixed levels of censoring (10%, 20%, 50%), to estimate the parameter of interest and to compare the proposed method with the corresponding method that assumes independent censoring. They concluded that their weighted estimator outperformed the 'naive' unweighted estimator. With an example dataset, WQC and MKLB compared the results of the application of their method with the 'naive' corresponding method.

Although the finite sample properties of the proposed estimators had been studied for each of those methods and the advantages of MKLB approach over the corresponding method assuming independent censoring have been established, there has been no systematic study to compare these two fairly recent methods. Since each method investigates dependent censoring through distinct mechanism, it would be of interest to evaluate the relative performance of the two estimators under various scenarios for the dependent censoring. Due to the complexity of the data structure and the estimation approaches, an analytic comparison in general seems very difficult, if not impossible, to

obtain. Hence, we conducted simulation studies to compare these two approaches for the estimation of covariate effects from recurrent time-to-event data in the presence of dependent censoring.

We summarized the WQC and MKLB approaches in Section 5.2. In Section 5.3, we outline the simulation framework and in Section 5.4 we present the results of the simulation. In Section 5.5, we provide an illustration with recurrent diarrheal data from the community trial on vitamin A supplementation. The conclusions appear in Section 5.6.

5.2 Approaches for Recurrent Event Data with Dependent Censoring

Let $N(t)$ be the number of recurrent events at or before t, $t \geq 0$, and suppose that the occurrence rate of recurrent events in the interval $[0, \tau]$ is of interest, where τ refer to the end time of the study. Let C denote the follow-up or censoring time and $Y(t) = I(C \geq t)$ be the at-risk indicator. Thus, in order to explore the association between the covariates $\mathbf{Z}(t)$ and N(.), consider the following rates model:

$$E\{dN(t)|\mathbf{Z}(t)\} = Y(t)d\mu_0(t)\exp\{\beta'\mathbf{Z}(t)\}$$

where $dN(t)$ denotes the number of events in the small time interval $[t, t+dt]$, β is a $p \times 1$ vector of fixed regression parameters and $d\mu_0(t)$ is an unspecified baseline rate function.

WQC proposed to model the occurrence of recurrent events by a subject-specific nonstationary Poisson process via a latent variable, allowing the censoring mechanism be possibly informative. The distribution of both the censoring and latent variables are treated as nuisance parameters. They assume that there exists a nonnegative valued latent variable U so that, conditioning on $(\mathbf{Z}(t), u)$, $N(t)$ is a nonstationary Poisson process with intensity function $u\lambda_0(t)\exp\{\beta'\mathbf{Z}(t)\}$, where the baseline intensity $\lambda_0(t)$ is a continuous function. The latent variable satisfies $E[U|\mathbf{Z}(t)] = 1$. This assumption

implies the marginal proportional rate function defined above. The second assumption is that conditioning on $(\mathbf{Z}(t), u)$, N(.) is independent of C.

Considering a conditional likelihood that involves only the shape function F and does not require information on the unobserved $\{u_i\}$, WQC defined a class of estimating equations for $\gamma' = (\ln \beta_0, \beta')$ as

$$n^{-1} \sum_{i=1}^{n} w_i \bar{\mathbf{Z}}_i(t)' (k_i \hat{F}^{-1}(C_i) - e^{\gamma' \bar{\mathbf{Z}}_i(t)}) = 0,$$

where $\bar{\mathbf{Z}}_i(t) = (1, \mathbf{Z}_i(t))$, $\beta_0 = \Lambda_0(\tau)$, $\Lambda_0(\tau)$ denotes the baseline cumulative rate function, k_i is the number of recurrences for subject i, and w_i is a weight function depending on $(\mathbf{Z}_i(t), \gamma, F)$. Under regularity conditions, the estimator $\hat{F}(t)$ is known to have a simple product-limit representation, such that $\hat{F}(t) = \prod_{s_{(l)} > t}(1 - d_{(l)}/N_{(l)})$, where $\{s_{(l)}\}$ are the ordered and distinct values of the event times $\{T_{ij}\}$, $\{d_{(l)}\}$ is the number of events occurring at $s_{(l)}$, and $\{N_{(l)}\}$ is the total number of events with event time and censoring time satisfying $T_{ij} \leq s_{(l)} \leq C_i$.

Wang, Qin and Chiang (2001) showed that the solution of this class of estimating equations has the property that $\sqrt{n}(\hat{\gamma} - \gamma)$ converges weakly to a multivariate normal distribution with zero mean and covariance matrix which can be consistently estimated if the marginal rate model is correctly specified.

The MKLB method uses inverse probability of censoring weighted(IPCW) estimators for the regression parameters in the proportional rates model defined previously, where $\mathbf{Z}(t)$ is a function of $\bar{\mathbf{Z}}^*(t)$, and $\bar{\mathbf{Z}}^*(t) \subset \mathbf{Z}^*(t)$, consisting of part of the covariate process $\mathbf{Z}^*(t)$. The authors proposed estimating equations for the parameter of interest in this general model by using IPCW mapping (Robbins and Rotnitzky, 1992), for which the main idea is to map full data estimating functions into observed data estimating functions. The class of all full data estimating function for the proportional rates model is given by $D_h^* = \int h^*\{t, \bar{\mathbf{Z}}^*(t)\} dM_r(t)$, where $dM_r(t) \equiv dN(t) - E\{dN(t) | \bar{\mathbf{Z}}^*(t-)\}$ and h^* is a user-defined function (van der Laan and Robins, 2002).

A choice of IPCW estimating function is then given by:

$$U_G(\mathbf{E}|D_h^*) = \int_0^\tau h^*\left\{t, \bar{\mathbf{Z}}^*(t-)\right\} \frac{dM_{\beta,\lambda_0}(t)Y(t)}{\bar{G}(t|\mathbf{V})},$$

where $\bar{G}(t|\mathbf{V}) = P(C > t|\mathbf{V})$ and $\mathbf{V} = \bar{V}(\tau) = \{N(\tau), \mathbf{Z}^*(\tau)\}$, stands for everything that can be observed on a randomly selected subject in the interval $(0, \tau]$ if the subject is not subject to censoring. Often the full data is not observed but their censored version. Denote the observed data random variable by $\mathbf{E}$=$\{min(\tau, C), \Delta^* = I(\tau < C), \bar{V}(\tau \wedge C)\}$.

This methodology requires a model for the censoring mechanism, which can be given by:

$$\lambda_c(t|\bar{V}(t-)) = Y_c(t)\lambda_{0,c}(t)\exp\{\beta_c\xi_c(t)\},$$

where $Y_c(t)$ is the at-risk indicator for censoring, $\lambda_{0,c}(t)$ is an unspecified baseline hazard and $\xi_c(t)$ is a known function of $\bar{V}(t-)$.

Note that $U_G(.|D_h^*)$ satisfies $E\{U_G(\mathbf{E}|D_h^*)|\mathbf{V}\} = D_h^*(\mathbf{V}|\beta, \lambda_0)$ under the assumption that $P(C > \tau|\mathbf{V}) > \delta > 0$, for some $\delta > 0$ and hence it yields consistent estimators in the presence of dependent censoring.

Applying time-dependent weighting to the full data estimating equation yields the following observed data estimating equation:

$$\begin{aligned}\mathbf{U}_G(\mathbf{E}|D_h^*) &= \int_0^\tau \left(\mathbf{Z}(t) - \frac{E[I(C > t)\bar{G}(t|\mathbf{V})^{-1}\mathbf{Z}(t)\bar{G}\{t|\bar{\mathbf{Z}}^*(t-)\}Y(t)\exp\{\beta\mathbf{Z}(t)\}]}{E[I(C > t)\bar{G}(t|\mathbf{V})^{-1}\bar{G}\{t|\bar{\mathbf{Z}}^*(t-)\}Y(t)\exp\{\beta\mathbf{Z}(t)\}]}\right) \\ &\quad \times \frac{\bar{G}\{t|\bar{\mathbf{Z}}^*(t-)\}Y(t)dM_{\beta,\lambda_0}(t)}{\bar{G}(t|\mathbf{V})}.\end{aligned}$$

Given estimators $\hat{h}^*$, $\hat{\bar{G}}$ and $\hat{\lambda}_0$ of h^*, $\bar{G}$ and λ_0, an estimator for β can be obtained by solving the estimating equation:

$$\sum_{i=1}^n \mathbf{U}_G\{\mathbf{E}_i|\hat{\bar{G}}, \hat{D}_h^*(.|\beta, \hat{\lambda}_0)\} = \mathbf{0}$$

Note that $\bar{G}$ is estimated by fitting the multiplicative intensity model for the censoring process. The estimate for h^* is then obtained by substituting $\hat{\bar{G}}$ for $\bar{G}$ and estimating the expectations empirically. The authors mention that one of the strengths of the method is that it can be easily implemented by adapting standard routines available in statistical software packages.

Both approaches characterizes the rate of the counting process under the marginal rates model, allowing arbitrary dependence structures among recurrent events. However, the two approaches differ in their ways to adjusting for dependent censoring. WQC introduces a latent variable to handle the informative or dependent censoring, while MKLB deal with the problem of informative censoring by modelling the censoring time using observable covariate information.

5.3 Simulation Framework

Consider a clinical trial where each subject is randomly assigned to a treatment arm of interest. Let $N(t) = \sum_k I(T_k \leq t)$ be the recurrent events counting process of interest while Z denotes the treatment variable and W a baseline covariate. Suppose that the goal of this study is to estimate the effect of the treatment. Assuming a proportional rates model, the parameter of interest is the regression coefficient β in the following model:

$$d\mu(t|Z) = d\mu_0(t)\exp(\beta Z)$$

In many applications censoring could be caused by informative dropouts or failure events and it is unrealistic to assume independence between the censoring mechanism and the recurrent event process. If C is not independent of T given Z, then the estimator for β will be inconsistent when using an *ad hoc* estimation method by fitting a marginal rates model for the right-censored data on T ignoring any information beyond Z.

In order to mimic such study and compare the two aforementioned methodologies to handle dependent censoring, T is generated using the following intensity function

$\lambda_T(t|Z,W,u) = u\lambda_{0,T}(t)\exp(\beta_0 Z + \gamma_0 W)$. Conditioning on (Z,W,u), the censoring time C is generated from $\lambda_C(t|Z,W,u) = u\lambda_{0,C}(t)\exp(\tilde{\beta}_0 Z + \tilde{\gamma}_0 W)$.

We generated independent Z from Bernoulli distribution and W from the following distributions: Uniform, Bernoulli and Normal. The failure indicator Δ_i is defined as $\Delta_i = I(T_i \leq C_i)$. We generated independent u_i (i=1,...,n) from gamma distribution, with mean 1 and variance σ^2. Large values of σ^2 reflect greater heterogeneity between subjects and a stronger association between events from the same subject.

The focus of this simulation study is on the performance of WQC and MKLB approaches in the estimation of β for various combinations of σ^2, sample size, treatment effect and baseline covariate effect. Each simulated data set consists of information from n independent subjects, with $N_i(t)$, i=1...,n, $t \in [0,\tau]$ denoting the number of recurrent events for the ith subject.

We generated 1,000 samples for each configuration of simulation parameters. We used the sample bias and sample variances to measure, respectively, the accuracy and efficiency of regression parameter estimates from the two approaches. The mean squared errors were also computed using the sample bias and variances. The sample bias and sample variance are defined, respectively, as the average bias and the variance from the 1,000 random samples. Let $\hat{\beta}_i$ be the estimate of the ith random sample, then:

$bias = \frac{\sum \hat{\beta}_i}{1{,}000} - \beta$, Sample variance=$\frac{\sum(\hat{\beta}_i - \bar{\beta})^2}{1{,}000-1}$, where $\bar{\beta} = \frac{1}{1{,}000}\sum \hat{\beta}_i$.

Note that the parameter of interest β is not the same as the β_0 which is used to generate the data through a conditional model. According to Miloslavsky et al (2004), we obtain a good estimate of the true parameter β by generating a large number of observations (e.g. N=100,000) from the data-generating distribution and fitting the marginal model using the full data(T,Z,W). This estimate corresponds to the minimizer of the Kullback-Leibler projection of the true data-generating distribution of the model of interest.

The simulation study was conducted in R v.1.9.1 software. The standard coxph( ) routine in R, providing the appropriate weights, was used for the MKLB approach. We used the crf R library, developed by WQC, in order to fit their model. This library is available upon request to the authors of WQC approach.

TABLE 5.1: Simulation results on bias, empirical standard error and mean squared error of the three estimators for the regression parameter β based on 1,000 replicates (with $\beta_0 = -1.2$, $\tilde{\beta}_0 = 1$) and $W \sim$ Uniform(3,4)

			MKLB Method			WQC Method			Indep.cens. Method		
σ^2	$(\gamma_0, \tilde{\gamma}_0)$	n	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
0	(0,0)	500	0.0038	0.0642	0.0041	0.0064	0.1326	0.0176	0.0031	0.0641	0.0041
		200	0.0064	0.1040	0.0109	0.0165	0.1969	0.0390	0.0091	0.1037	0.0108
	(8,5)	500	0.0046	0.1733	0.0301	0.0079	0.1908	0.0365	0.1912	0.1659	0.0641
		200	0.0580	0.2819	0.0828	0.0748	0.2977	0.0942	0.1388	0.2689	0.0916
1	(0,0)	500	0.2347	0.1103	0.0673	0.0215	0.1654	0.0278	0.2339	0.1101	0.0668
		200	0.2357	0.1786	0.0875	0.0422	0.2256	0.0527	0.2345	0.1776	0.0865
	(8,5)	500	0.1052	0.2398	0.0686	0.0641	0.2621	0.0728	0.1730	0.2288	0.0896
		200	0.0498	0.3317	0.1125	0.1344	0.3595	0.1473	0.1356	0.3182	0.1196
4	(0,0)	500	0.4613	0.1585	0.2379	0.0811	0.1942	0.0443	0.4612	0.1583	0.2378
		200	0.3554	0.2460	0.1868	0.1683	0.2509	0.0913	0.3559	0.2459	0.1871
	(8,5)	500	0.2362	0.2958	0.1433	0.1743	0.3312	0.1401	0.2692	0.2843	0.1533
		200	0.1600	0.4334	0.2134	0.2776	0.5020	0.3291	0.1972	0.4155	0.2115

5.4 Simulation Results

For the results summarized in Table 5.1, the parameters of the data-generating distributions were set as follows: $\beta_0 = -1.2$, $\gamma_0 = 0$ and 8, $\tilde{\beta}_0 = 1$, $\tilde{\gamma}_0 = 0$ and 5, $\sigma^2 = 0$, 1 and 4, $\tau = 4$ months, n=200 and 500, $Z \sim$ Bernoulli(0.5) and $W \sim$ Uniform(3,4). The average number of events per subject is about 3.6.

Under the scenario of independent censoring ($\sigma^2 = 0, \gamma_0 = 0, \tilde{\gamma}_0 = 0$), the use of both methods leads to approximately unbiased estimates. In addition, the results obtained from the MKLB method are similar to those for proportional rates model without dependent censoring (Lin et al, 2000). When censoring is dependent through covariates ($\sigma^2 = 0, \gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$), modeling the censoring mechanism with proper covariate information using IPCW estimators in the MKLB method is approximately unbiased while the estimate from the proportional rates model ignoring dependent censoring is biased (bias=0.1912 for n=500). In such scenario, the MKLB estimator is less biased and more precise than the WQC estimator (Table 5.1). However WQC estimator is also approximately unbiased.

Wang et al (2001) use a latent variable to characterize the heterogeneity among subjects and assume that the latent variable u_i is the only factor that explains the heterogeneity from different subjects (besides Z_i). It is evident from Table 5.1 that in this case ($\sigma^2 = 4, \gamma_0 = 0, \tilde{\gamma}_0 = 0$), the estimator from WQC method (bias=0.0811 for n=500) are much less biased than that obtained from MKLB method (bias=0.4613 for n=500). Similar patterns were observed when the variability of the latent variable was reduced($\sigma^2 = 1, \gamma_0 = 0, \tilde{\gamma}_0 = 0$). The bias for the MKLB method is smaller for smaller σ^2. Same was observed for WQC method. At the same time, when considering that the censoring mechanism depends not only on the observed baseline covariates (W) but also on unmeasured factors (u), both estimators become biased (bias=0.2362 and 0.1743 for MKLB and WQC methods, respectively, for n=500).

The empirical standard errors (ESE) of MKLB method were consistently smaller than those obtained using WQC method. We also compared these methods through the use of the mean squared error (MSE) as the comparison criterion. Note that for

TABLE 5.2: Simulation results on bias, empirical standard errors and mean squared errors of the three estimators for the regression parameter β based on 1,000 replicates (with $\beta_0 = -1.2$, $\tilde{\beta}_0 = 1$) and $W \sim$ Bernoulli(0.5)

			MKLB Method			WQC Method			Indep.cens. Method		
σ^2	$(\gamma_0, \tilde{\gamma}_0)$	n	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
0	(0,0)	500	0.0027	0.0640	0.0041	0.0104	0.1160	0.0136	0.0017	0.0639	0.0041
		200	0.0011	0.0979	0.0096	0.0033	0.1698	0.0288	0.0021	0.0972	0.0094
	(8,5)	500	0.0398	0.1152	0.0149	0.0194	0.1586	0.0255	0.2260	0.1306	0.0681
		200	0.0361	0.1808	0.0340	0.0203	0.2290	0.0529	0.2211	0.2053	0.0911
1	(0,0)	500	0.0398	0.1106	0.0138	0.0057	0.1248	0.0156	0.0397	0.1105	0.0138
		200	0.0496	0.1628	0.0280	0.0001	0.1888	0.0356	0.0496	0.1628	0.0290
	(8,5)	500	0.1087	0.1709	0.0410	0.0371	0.1876	0.0366	0.1454	0.1701	0.0501
		200	0.0947	0.2590	0.0760	0.0342	0.2767	0.0777	0.1303	0.2600	0.0846
4	(0,0)	500	0.1117	0.1926	0.0496	0.0047	0.2119	0.0449	0.1119	0.1924	0.0495
		200	0.0814	0.2809	0.0856	0.0299	0.2996	0.0907	0.0819	0.2808	0.0856
	(8,5)	500	0.1334	0.2720	0.0918	0.0252	0.2881	0.0837	0.1459	0.2699	0.0941
		200	0.1055	0.3910	0.1641	0.0042	0.4084	0.1668	0.1168	0.3884	0.1645

the results presented in Table 5.1, the smallest MSE was generally observed for the corresponding method with the smallest bias.

When $W \sim$ Bernoulli(0.5) and $W \sim$ Normal(0,1) (Tables 5.2 and 5.3, respectively), the magnitude of the bias was generally reduced for all scenarios compared to the results presented in Table 5.1. Regardless of the distribution associated with W, the smallest bias and ESE for all methods were generally observed under the scenario of independent censoring ($\sigma^2 = 0, \gamma_0 = 0, \tilde{\gamma}_0 = 0$) and the MKLB method has the smallest MSE when censoring is dependent through covariate W ($\sigma^2 = 0, \gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$).

The WQC method outperforms the MKLB method in terms of bias when the dependence between event and censoring times is introduced only through a latent variable ($\sigma^2 = 1$ or 4,$\gamma_0 = 0, \tilde{\gamma}_0 = 0$). Due to the general reduced magnitude of the bias when $W \sim$ Bernoulli(0.5) and $W \sim$ Normal(0,1), the values of MSE in Tables 5.2 and Table 5.3 are strongly influenced by ESE, which are consistently smaller for MKLB method. Hence, in those scenarios the MSE will be mostly driven by the efficiency instead of by the bias of the estimates.

The worst performance was observed when the censoring mechanism depends on both the observed baseline covariate (W) and on unmeasured factors (u) ($\sigma^2 = 1$ or 4, $\gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$). For all parameter configurations considered in these simulation studies, the sampling variances increase as the sample size decreases from 500 to 200.

To compare the effect of the relative magnitude of W on the estimation process regardless of the probability distribution associated with it, we display in Table 5.4 the results of a simulation study considering $W \sim$ Uniform(0,1). As it was observed for the simulation studies with $W \sim$ Bernoulli(0.5) and $W \sim$ Normal(0,1), the bias magnitudes were generally reduced compared to when $W \sim$ Uniform(3,4). The WQC approach again has the least bias in the presence of a latent variable and without any effect of W on the event occurrence ($\sigma^2 = 1$ or 4, $\gamma_0 = 0, \tilde{\gamma}_0 = 0$). However, when introducing the dependent censoring through covariate W ($\gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$), the bias for the WQC approach was slightly smaller than that for MKLB approach for $\sigma^2 = 0$ while MKLB approach outperforms WQC approach in terms of bias when $\sigma^2 \neq 0$. Such results are different from those obtained when $W \sim$ Uniform(3,4) and somewhat

TABLE 5.3: Simulation results on bias, empirical standard errors and mean squared errors of the three estimators for the regression parameter β based on 1,000 replicates (with $\beta_0 = -1.2$, $\tilde{\beta}_0 = 1$) and $W \sim$ Normal(0,1) truncated at (-1,1)

			MKLB Method			WQC Method			Indep.cens. Method		
σ^2	$(\gamma_0, \tilde{\gamma}_0)$	n	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
0	(0,0)	500	0.0049	0.0634	0.0040	0.0088	0.1408	0.0199	0.0060	0.0633	0.0040
		200	0.0048	0.1008	0.0102	0.0011	0.1834	0.0336	0.0065	0.1005	0.0102
	(8,5)	500	0.0046	0.1694	0.0287	0.0180	0.1757	0.0312	0.1041	0.1780	0.0425
		200	0.0373	0.2737	0.0763	0.0255	0.2890	0.0842	0.0542	0.2794	0.0810
1	(0,0)	500	0.0502	0.1059	0.0137	0.0046	0.1275	0.0163	0.0502	0.1058	0.0137
		200	0.0355	0.1634	0.0280	0.0098	0.1809	0.0328	0.0355	0.1633	0.0279
	(8,5)	500	0.0233	0.4797	0.2307	0.1398	0.4817	0.2516	0.0788	0.4582	0.2162
		200	0.1323	0.6479	0.4373	0.2557	0.6383	0.4728	0.0629	0.6293	0.3999
4	(0,0)	500	0.0647	0.1814	0.0371	0.0433	0.2015	0.0425	0.0647	0.1815	0.0371
		200	0.0916	0.2879	0.0913	0.0173	0.3036	0.0925	0.0937	0.2877	0.0915
	(8,5)	500	0.0956	0.4106	0.1777	0.1013	0.3943	0.1657	0.1266	0.4054	0.1803
		200	0.1055	0.3911	0.1641	0.0042	0.4084	0.1668	0.1168	0.3884	0.1645

similar to those obtained when $W \sim$ Normal(0,1).

In summary, all methods are approximatelly unbiased for the scenario of independent censoring ($\sigma^2 = 0, \gamma_0 = 0, \tilde{\gamma}_0 = 0$) and the WQC approach produces the least efficient estimates. When the only source of dependent or informative censoring is known to be due to the covariates ($\sigma^2 = 0, \gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$), which can be properly modelled through the censoring mechanism, the MKLB method generally yield the most accurate and efficient estimates compared to WQC method. Particularly in such scenario the MKLB estimates are much less biased than those obtained by fitting a marginal rates model under the assumption of independent censoring regardless of the relative magnitude and distribution of W. Nevertheless, when the heterogeneity among subjects was introduced only through a latent variable ($\sigma^2 \neq 0, \gamma_0 = 0, \tilde{\gamma}_0 = 0$), WQC approach always outperforms MKLB approach in terms of accuracy for the configurations studied here.

On the other hand, when both covariate (W) and latent variable (u) were used to introduce dependent censoring on the event occurrence ($\sigma^2 \neq 0, \gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$), the results are not consistent across the parameter configurations considered in this paper. In those situations, the accuracy and efficiency of the estimates seems to vary for distinct relative magnitudes as well as probability distributions associated to W.

5.5 An Example: Modelling Times to Recurrent Diarrhea in Children

In this Section we apply the aforementioned methods to recurrent diarrhea data to illustrate the modelling process. We used data from 1,191 children aged 6-48 months at baseline, who participated in a randomized community trial conducted in Brazil between 1990 and 1991 to evaluate the effect of high dosages of vitamin A supplementation on the occurrence of recurrent diarrheal episodes. The complete study was described in Barreto et al (1994). For the analysis presented here, we consider the data available from the first treatment cycle, i.e, between the first and second dosages of vitamin A. During this period, the mean number of episodes of diarrhea was 2.526

TABLE 5.4: Simulation results on bias, empirical standard errors and mean squared errors of the three estimators for the regression parameter β based on 1,000 replicates (with $\beta_0 = -1.2$, $\tilde{\beta}_0 = 1$) and $W \sim$ Uniform(0,1)

			MKLB Method			WQC Method			Indep.cens. Method		
σ^2	$(\gamma_0, \tilde{\gamma}_0)$	n	Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
0	(0,0)	500	0.0035	0.0649	0.0042	0.0073	0.1246	0.0156	0.0001	0.0976	0.0095
	(8,5)	500	0.0173	0.1764	0.0314	0.0136	0.1836	0.0339	0.0383	0.1736	0.0316
1	(0,0)	500	0.0361	0.1058	0.0125	0.0131	0.1278	0.0165	0.0361	0.1059	0.0125
	(8,5)	500	0.0379	0.2365	0.0574	0.0593	0.2609	0.0716	0.0987	0.2283	0.0619
4	(0,0)	500	0.1178	0.1923	0.0509	0.0099	0.2067	0.0428	0.1180	0.1922	0.0509
	(8,5)	500	0.0266	0.3391	0.1157	0.1284	0.3366	0.1509	0.0471	0.3346	0.1142

(sd=2.41, range=0-15). The covariates include demographic, economic and health indicators. Therefore, all 1,191 subjects included in this analysis have complete information for all indicators. We consider the following covariates for modelling diarrhea occurrence: age (in months, at baseline), sex, treatment group (placebo or vitamin A) and an indicator of existence of toilet (TOILET) in the household. To capture their health status we consider as covariates weight-for-age Z-score (WAZ) and previous occurrence of measles.Among these children, 26.4% lived in houses that do not have toilets and 89.3% had measles previously.

The dependent censoring could have been introduced in this study if children who were at higher risk of having recurrent diarrheal episodes withdrawn from the study earlier. Another form of dependent censoring could have been introduced due to terminal events, such as death. However, the few death cases occurred during this study was equally distributed among the two treatment groups and were not associated to diarrhea occurrence. Thus we explore the possibility that dependent or informative censoring had occurred in the vitamin A study by estimating the parameters for the model of interest through the use of the methodologies proposed by WQC and MKLB. We compared the results from fitting WQC and MKLB methods with those obtained from fitting a standard marginal rates model. Assuming that the censoring mechanism is independent of the counting process of interest given the covariates that we are conditioning on, we fit the standard marginal rates model and go through a model selection procedure. The estimated coefficients for the final model is given in Table 5.5.

TABLE 5.5: Estimated coefficients for the marginal rates model of diarrhea occurrence assuming independent censoring

Model	$\hat{\beta}$	Estimated robust SE($\hat{\beta}$)	p-value
TRT	-0.136	0.0556	0.0150
AGE	-0.030	0.0024	< 0.0001
TOILET	0.254	0.0622	< 0.0001
WAZ	-0.050	0.0181	0.0057

After that we estimate the parameters using WQC and MKLB methods. The first

step in applying MKLB approach is to obtain a 'good' model for the censoring mechanism. For the selection of such model, we initially considered all the covariates that we also considered when fitting the marginal rates model of the diarrhea occurrence. The estimated coefficient for the censoring mechanism model is presented in Table 5.6. The only covariate important for the censoring mechanim was WAZ.

TABLE 5.6: Estimated coefficients for the censoring marginal rates model

Model	$\hat{\beta}$	Estimated robust SE($\hat{\beta}$)	p-value
WAZ	-0.059	0.0166	0.0004

The next step is to estimate the weights by using the estimated censoring survival probability that is obtained by selecting a model for the censoring mechanism. The regression coefficients that were estimated by the IPCW estimating function for MKLB approach are given in Table 5.7. We also estimated those parameters considering the method proposed by WQC. The standard errors for both methods were estimated using bootstrap. The corresponding estimates are also presented in Table 5.7.

TABLE 5.7: Estimated coefficients for the marginal rates model of diarrhea occurrence using WQC and MKLB approaches

Variables	MKLB			WQC		
	Parameter estimate	Standard error	p-value	Parameter estimate	Standard error	p-value
TRT	-0.137	0.0568	0.0159	-0.131	0.0516	0.0111
AGE	-0.030	0.0027	< 0.0001	-0.025	0.0021	< 0.0001
TOILET	0.253	0.0630	< 0.0001	0.210	0.0558	0.0002
WAZ	-0.051	0.0239	0.0329	-0.042	0.0166	0.0114

By comparing Tables 5.5 and 5.7, we note that the estimated coefficients and standard errors do not change noticeably when we employ the approaches that take into account the dependent censoring. On the basis of the results from the MKLB approach, the covariate that was important for the censoring mechanism was already included in

the marginal rates model for diarrhea occurrence, which will lead to the assumption of independent censoring conditional on the covariates that are included in the model of interest. The results from WQC approach do not point out for any other source of dependent censoring in this data either.

Both models lead to the same clinical conclusions. TRT has a strong effect on the rate of diarrhea occurrence (RR=0.87). Based on these results, the rate of diarrhea occurrence in children receiving vitamin A supplementation is 13% lower than the corresponding rate in children in the placebo group. The increase in age and in weight-for-age Z-score also contributes for a significant reduction on the rate of diarrhea occurrence. On the contrary, the existence of toilet in the house leads to an increase of 29% on the rate of diarrhea, which could be associated to poor hygiene practices in this community.

5.6 Conclusion

We compared two approaches (WQC and MKLB) for the estimation of covariate effects for recurrent time-to-event data and found that they produce approximately unbiased estimates when the dependent or informative censoring is not present. The variances of the parameter estimates from the two approaches increase with decreasing sample size, as expected. Generally, the empirical standard errors from WQC approach are larger than those from MKLB approach. According to Wang et al (2001), this later approach is expected to achieve optimal estimation efficiency at the price of modelling the censoring mechanism with proper covariate information. Biased results were found when the informative or dependent censoring was introduced simultaneously by a covariate and a latent variable.

Overall, MKLB method outperforms the usual marginal rates model in terms of bias and accuracy when the informative censoring was introducing through a covariate ($\sigma^2 = 0,\ \gamma_0 \neq 0, \tilde{\gamma}_0 \neq 0$). Similar pattern was observed through the comparison of WQC model and marginal rates model when the informative censoring was introduced

via a latent variable ($\sigma^2 = 1$ or 4, $\gamma_0 = 0, \tilde{\gamma}_0 = 0$). Further research is still needed for more complex situations, particularly when there exists more than one source of informative censoring and the censoring mechanism can not be properly modelled with available covariate information.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

In many situations it is of interest to explore the functional form of the relationship between covariates and failure time. Furthermore, it may be important to allow the coefficients to change over time. Even though several alternatives of the Cox model have been made to allow the effects to change over time, no time-varying effects model had been proposed for the recurrent time-to-event settings. To address this issue, we proposed methods to estimate time-varying coefficients using B-splines in the marginal rates model to analyze recurrent time-to-event data. Two approaches were used for the estimation of time-varying coefficients: regression and penalized splines.

First, we considered the introduction of splines with small number of knots in the marginal rates model and applied standard procedures to obtain the estimates for the time-varying effects. The results of simulations indicate that our estimates of the rate ratios are approximately unbiased and the variance estimator performs well, unless the data are very scarce. Our method considered that the number of knots is held fixed as the sample size $n \to \infty$, such that the B-spline basis is chosen a priori. Further study might propose alternative estimators for more complex cases, in which is not known a priori which among the effects of several correlated variables are time-varying and which are constant. Even though we discussed posteriori model selection criteria, such as GCV and AIC, which can be helpful to find a reasonable trade-off between model parsimony and the risk of overfitting bias, quantifying the variance inflation due to the

use of such criteria is a potential area for future research.

We also proposed a method that uses B-splines with small to moderate number of knots and that estimates the time-varying effects in the marginal rates model with estimation based on penalized pseudo-partial likelihood. The penalty functions do not allow the estimated functions to fluctuate rapidly, resulting in more stable estimates. Simulation results demonstrate that the asymptotic inference procedures are reliable when adequately large sample sizes are used. The small sample properties of the proposed estimators may be improved by incorporating different choices of the penalty function and can be explored.

Simulation studies were also conducted to compare two recently proposed methods for recurrent event data that take into account the presence of dependent or informative censoring. Results pointed out that the MKLB estimator has better performance when the censoring mechanism depends on observed covariates that are included in the model of interest while the WQC method may handle situations where the censoring is caused by some unmeasured or latent factor. When censoring is related to both the observed covariates and some unmeasured factors, neither method work well. New approaches are needed for such more complex situations.

REFERENCES

Abdeljaber, M.H., Monto, A.S., Tilden, R.L., Schork, A. and Tarwotjo, I. (1991). The Impact of Vitamin A Supplementation on Morbidity: A Randomized Community Intervention Trial, *American Journal of Public Health* 1654-1656.

Abrahamowicz, M., MacKenzie, T. and Esdaile, J.M. (1996). Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing with Application in Lupus Nephritis, *The Journal of the American Statistical Association* 91, 1432-1439.

Andersen, P. K. and Gill (1982). Cox's Regression Model for Counting Processes: A Large Sample Study, *The Annals of Statistics* 10, 110-1120.

Barreto, M.L., Santos, L.M.P., Assis, A.M.O., Araujo, M.P.N., Farenzena, G.G., Santos, P.A.B., Fiaccone, R.L.(1994). Effect of vitamin A supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in Brazil, *Lancet* 344, 228-231.

Berhane, K. and Weissfeld, L.A. (2003). Inference in Spline-Based Models for Multiple Time-to-Event Data, with Applications to a Breast Cancer Prevention Trial, *Biometrics* 59, 859-868.

Bhandari, N., Bhan, N.K., Sazawal, S. (1994). Impact of massive dose of vitamin A given to preschool children with acute diarrhoea on subsequent respiratory and diarrhoeal morbidity, *BMJ* 309, 1404-1407.

Biswas, R., Biswas, A.B., Manna, B, Bhattacharya, S.K., Dey, R., Sarkar, S. (1994). Effect of vitamin A supplementation on diarrhoea and acute lower respiratory tract infection in children,*European Journal of Epidemiology* 10, 57-61.

Buja, A., Hastie, T., Tibshirani, R. (1989). Linear Smoothers and Additive Models, *The Annals of Statistics* 17, 453-555.

Byers, K.E., Guerrant, R.L. and Farr, B.M.(2001). *In: Epidemiologic Methods for the Study of Infectious Diseases*, Thomas, J.C, Weber, D.J. (eds), Oxford University Press, New York: Oxford, 228-248.

Cai, J. and Schaubel, D. (2004). Analysis of Recurrent Event Data, *Handbook of Statistics* 23, 603-623.

Cai, Z. and Sun, Y. (2003). Local Linear Estimation for Time-Dependent Coefficients

in Cox's Regression Models, *Scandinavian Journal of Statistics* 30, 93-111.

Chiang, S.-H. (1968). *Regression Analysis for recurrent event data*, Doctoral Dissertation, Johns Hopkins University: Department of Biostatistics.

Chowdhury S, Kumar R, Ganguly NK, Kumar L, Walia BN. (2002). Effect of vitamin A supplementation on childhood morbidity and mortality, *Indian J Med Sci* 56(6), 259-64.

Clayton, D. and Cuzick, J. (1985). Multivariate Generalizations of the Proportional Hazards Model, *Journal of Royal Statistics Society - Series A* 148, Part 2, 82-117.

Cook, R.J. and Lawless, J.F. (1997). Marginal Analysis of Recurrent Events and a Terminating Event, *Statistics in Medicine* 16, 911-924.

Cook, R.J. and Lawless, J.F. (2002). Analysis of Repeated Events, *Statistical Methods in Medical Research* 11, 141-166.

Cox, D.R. (1972). Regression Models and life-tables(with discussion), *Journal of Royal Statistics Society - Series B* 34, 182-220.

Cox, D.R. (1975). Partial likelihood, *Biometrika* 62, 269-276.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*, London: Chapman and Hall.

De Boor, C. (2001). *A Practical Guide to Splines*, Revised edition, New York: Springer.

Dibley MJ, Sadjimin T, Kjolhede CL, Moulton LH (1996). Vitamin A supplementation fails to reduce incidence of acute respiratory illness and diarrhea in preschool-age Indonesian children , *Journal of Nutrition* 126(2):434-42.

Duchateau, L., Janssen, P; Kezic, I. and Fortpied, C. (2003). Evolution of Recurrent Asthma Event Rate Over Time in Frailty Models, *Applied Statistics* 52, 355-363.

Eilers, P.H.C and Marx, B.D. (1996). Flexible Smoothing with B-Splines and Penalties, *Statistical Science* 11, 89-102.

Gamerman, D.(1991). Dynamic Bayesian methods for survival data, *Applied Statistics* 40, 63-79.

Gasser, T. and Muller,H-G. (1978).*In: Kernel Estimation of Regression Functions: Smoothing Techniques for Curve Estimation*, Gasser T, Rosenblatt M. (eds), Spring Lecture Notes in Mathematics, Berlin: Springer-Verlag, 23-68.

Ghosh, D. and Lin, D.Y. (2003). Semiparametric Analysis of Recurrent Event Data in the Presence of Dependent Censoring, *Biometrics* 59, 877-885.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Esteve, J, Gouvernet, J. and Faivre, J. (2003). A relative survival regression model using B-spline functions to model non-proportional hazards, *Statistics in Medicine* 22, 2767-2784.

Gray, R.J. (1992). Flexible Methods for Analyzing Survival Data using Splines, With Applications to Breast Cancer Prognosis, *Journal of the American Statistical Association* 87, 942-951.

Gray, R.J. (1994). Spline-Based Tests in Survival Analysis, *Biometrics* 50, 640-652.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models, *Journal of Royal Statistics Society - Series B* 55, Part 4, 757-796.

Hougaard, P. (1986). A class of Multivariate Failure Time Distributions, *Biometrika* 73, 671-678.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, New York: Springer-Verlag.

Huang, Y. and Chen, Y.Q. (2003). Marginal Regression of Gaps Between Recurrent Events, *Lifetime Data Analysis* 9, 293-303.

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

Kelly, P.J. and Lim, L. L-Y. (2000). Survival Analysis for Recurrent Event Data: An Application to Childhood Infectious Diseases, *Statistics in Medicine* 19, 13-33.

Klein, J.P. (1992). Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm, *Biometrics* 48, 795-806.

Lawless, J.F. and Nadeau, C. (1995). Some Simple Robust Methods for the Analysis of Recurrent Events, *Technometrics* 37, 158-168.

Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-Typed Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations, In: *Survival Analysis: State of the Art* 237-247.

Lie C, Ying C, Wang EL, Brun T, Geissler C. (1993).Impact of large-dose vitamin A supplementation on childhood diarrhoea, respiratory disease and growth, *European*

Journal of Clinical Nutrition,47(2), 88-96.

Lin, D.Y.(1994). Cox regression analysis of multivariate failure time data, *Statistics in Medicine* 15, 2233-2247.

Lin, D.Y.; Sun, W. and Ying, Z. (1999). Nonparametric Estimation of the Gap Time Distributions for Serial Events with Censored Data, *Biometrika* 86, 59-70.

Lin, D.Y. ; Wei, L.J.; Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events, *Journal of Royal Statistics Society - Series B* 62, Part 4, 711-730.

Lin, D.Y. and Ying, Z. (2001). Nonparametric Tests for the Gap Time Distributions of Serial Events Based on Censored Data, *Biometrics* 57, 369-375.

McGilchrist, C.A. and Aisbett, C.W. (1991). Regression with Frailty in Survival Analysis, *Biometrics* 47, 461-466.

Miloslavsky, M, Keles, S., van der Laan, M.J. and Butler, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring, *Journal of Royal Statistics Society - Series B* 66, Part 1, 239-257.

Morris, S.S., Cousens, S.N., Kirkwood, B.R., Arthur, P. and Ross, D.A. (1996). Is Prevalence of Diarrhea a Better Predictor of Subsequent Mortality and Weight Gain Than Diarrhea Incidence?, *American Journal of Epidemiology* 144(6), 583-588.

Moulton, L. H. and Dibley, M. J. (1997). Multivariate Time-to-Event Models for Studies of Recurrent Childhood Diseases, International Journal of Epidemiology 26(6), 1334-1339.

Murphy, S.A. and Sen, P.K. (1991). Time-dependent coefficients in a Cox-type regression model, *Stochastic Processes and Applications* 39, 153-180.

Nan, B; Lisabeth, L; Lin, X.; Harlow, S. (2003). A Varying-Coefficient Cox Model for the Effect of Age at a Marker Event on Age at Menopause, In: *The Working Paper Series of Department of Biostatistics. University of Michigan.*

O'Sullivan (1988). Nonparametric estimation of relative risk using splines and cross-validation, *SIAM Journal on Scientific and Statistical Computing* 9, 531-542.

Pepe, M.S. and Cai, J. (1993). Some graphical displays and Marginal Regression analysis for Recurrent Failure Times and Time Dependent Covariates, *Journal of American Statistical Association* 88, 811-820.

Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the Regression Analysis of Multivariate Failure Time Data, *Biometrika* 68, 373-379.

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/).

Rahman MM, Vermund SH, Wahed MA, Fuchs GJ, Baqui AH, Alvarez JO (2001). Simultaneous zinc and vitamin A supplementation in Bangladeshi children: randomised double blind controlled trial, *BMJ*, 323(7308), 314-8.

Ramakrishnan U, Latham MC, Abel R, Frongillo EA Jr. (1995). Vitamin A supplementation and morbidity among preschool children in south India, *American Journal of Clinical Nutrition* 61(6), 1295-303.

Robbins, J. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, In: *AIDS Epidemiology, Methodological Issues*. Boston: Birkhäuser, 297-331.

Rosenberg, PS. (1995). Hazard Function Estimation Using B-Splines, *Biometrics*, 51(3), 874-87.

Ross DA, Kirkwood BR, Binka FN, Arthur P, Dollimore N, Morris SS, Shier RP, Gyapong JO, Smith PG. (1995). 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*, 85(9), 1246-51.

Ross S.M. (1983) *Stochastic Processes*, New York: Wiley.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge: Cambridge University Press.

Schaubel, D. and Cai, J. (2004). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data, *Statistics in Medicine* 23, 1885-1900.

Sleeper, L.A. and Harrington, D.P. (1990). Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease, *Journal of the American Statistical Association* 85, 941-949.

Therneau, T. M. and Grambsch, P. M. (2001). *Modeling Survival Data*, New York: Springer.

Therneau, T. M., Grambsch, P. M. and Pankratz, V.S. (2003). Penalized Survival

Models and Frailty, *Journal of Computational and Graphical Statistics* 12(1), 156-175.

Tsiatis, A.A. (1981). A large sample study of Cox's regression model, *Annals of Statistics* 9, 93-108.

Valenta, Z. and Weissfeld, L. (2002). Estimation of the survival function for Gray's piecewise-constant time-varying coefficients model, *Statistics in Medicine*, 21, 717-27.

van der Laan, M.J. and Robins, J. M. (2002). *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer.

Victora, C.G., Bryce, J., Fontaine, O. and Monasch, R. (2000). Reducing Deaths from Diarrhea through Oral Rehydration Therapy, *Bulletin of the World Health Organization*, 78(10), 1246-1255.

UNICEF (2003). *Indicators for children and adolescents* (http://www.unicef.org.br). Accessed on Oct 2003.

Wang, M-C. and Chiang, C-T. (2002). Non-Parametric Methods for Recurrent Event Data With Informative and Non-Informative Censorings, *Statistics in Medicine* 21, 445-456.

Wang, M-C., Qin, J. and Chiang, C-T. (2001). Analyzing Recurrent Event Data With Informative Censoring, *Journal of the American Statistical Association* 96, 1057-1065.

Wei, L.J. and Glidden, D.V. (1997). An Overview of Statistical Methods for Multiple Failure Time Data in Clinical Trials, *Statistics in Medicine* 16, 833-839.

Wei, L.J, Lin, D.Y. and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions,*Journal of the American Statistical Association* 84, 1065-1073.

World Health Organization Department of Child and Adolescent Health and Development (1999). *The Evolution of Diarrheoeal and Acute Respiratory Disease Control: Achievements 1980-1995 in Research, Development, and Implementation.* Printed in France.