Multiple Time Scales and Longitudinal Measurements in Event History Analysis

Danardono

Statistical Studies No. 33
Department of Statistics
Umeå University 2005

Doctoral Dissertation

Department of Statistics
Umeå University
SE-901 87 Umeå, Sweden

Department of Public Health and Clinical Medicine,
Epidemiology and Public Health Sciences
Umeå University
SE-901 85 Umeå, Sweden

ISSN: 1100-8989
ISBN: 91-7305-812-2

Printed by Solfjädern Offset AB Umeå 2005

Abstract

A general time-to-event data analysis known as event history analysis is considered. The focus is on the analysis of time-to-event data using Cox's regression model when the time to the event may be measured from different origins giving several observable time scales and when longitudinal measurements are involved. For the multiple time scales problem, procedures to choose a basic time scale in Cox's regression model are proposed. The connections between piecewise constant hazards, time-dependent covariates and time-dependent strata in the dual time scales are discussed. For the longitudinal measurements problem, four methods known in the literature together with two proposed methods are compared. All quantitative comparisons are performed by means of simulations. Applications to the analysis of infant mortality, morbidity, and growth are provided.

Keywords and phrases: Cox regression, multiple events, proportional hazards, random effects, survival analysis, time-dependent covariates, time origin.

AMS subject classification: 62P10, 62N03.

To Leni, Fiyan and Lila

Acknowledgments

I would like to thank and express my deepest gratitude to:

Professor Göran Broström, my main supervisor, for his support and help during my studies and the writing of this thesis. I learned a lot from all our discussions during the last five years;

Dr. Hans Stenlund, my co-supervisor from the Department of Public Health and Clinical Medicine, Epidemiology and Public Health Sciences, for his support, comments and friendship;

Dr. Marie Lindkvist, who discussed the thesis manuscript in my *slutseminarium* and provided many valuable comments.

Professor Subanar, the dean of the Faculty of Mathematics and Natural Sciences, Gadjah Mada University, Indonesia, for his advice and support.

I would also like to thank the Community Health and Nutrition Research Laboratories (CHN-RL), Faculty of Medicine, Gadjah Mada University, for allowing me to use the surveillance data, and to Dr. Torbjörn Lind, for allowing me to use the ZINAK data.

I received financial support from STINT (Stiftelsen för internationalisering av högre utbildning och forskning - the Swedish foundation for international cooperation in research and higher education) during the initial stage of my studies at Umeå University, for my *licenciate* degree. Subsequently, I received financial support from Umeå University through the Department of Statistics and from the Department of Public Health and Clinical Medicine, Epidemiology and Public Health Sciences. To them I am very thankful.

Thanks to my many friends and colleagues who supported me during the life course of my studies. My warmest thanks to Birgitta Åström, for her friendship and endless assistance to me and my family. I also thank Anna Winkvist for her support, friendship and scientific discussions. To all Indonesian friends in Umeå, I say ”terima kasih banyak”.

Thanks (and goodbye...) to my ”old” classmates Jari’-san’, Maria, Marie; and to the ”younger”-mates, Mathias-ever-been-a-roommate, Ingeborg, Juke (thanks for your comments and corrections), Suad, Leake and Tea. Lycka till! ”Tack så mycket” to Birgitta Löfroth for your help and all my colleagues at the Department of Statistics, Umeå University.

To anyone else who, because of my limited memory, may have been omitted from being mentioned by name, I thank you for your assistance.

To Leni, Fiyan and Lila, my beloved family, thank you for supporting me and being here. I apologize, that my mind was often engaged with this thesis during dinner. I do not have enough words to thank you here. This thesis is dedicated to you.

I would also like to say something about my name. Many people asked me why I only have one name (one word). In Indonesia, where I come from, there is no requirement to have a family name. We have liberty to have our own name. I have one name, my wife and our children have three names (three words) each.

Finally, thanks for reading this thesis, at least this page...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Event history and longitudinal data

Event history and longitudinal data frequently arise in many scientific investigations. Important examples are in epidemiological surveillance and clinical trials. The nature of the data is that information on specific units or subjects are followed over time.

The term event history data possibly originated from sociology. Another applicable term is survival and duration data. Other popular terms for longitudinal data are repeated measurements, commonly used in biological or health sciences, and panel data, commonly used in the social sciences.

While, generally, event history and longitudinal data have many characteristics in common, their differences will be emphasized here. *Event history data* refers to *time-to-event* data, whereas *longitudinal data* refers mostly to *repeated measurements*. Two examples that will be used throughout this thesis are given below.

In 1994, an epidemiological surveillance was established in Purworejo district in Indonesia under the Community and Health Nutrition Laboratories (CHN-RL), Gadjah Mada University, Yogyakarta. Households were visited every 90-th day to record vital demographic events, morbidity events, nutritional status and utilization of health services (Wilopo and CHN-RL Team, 1997). The general aim of the surveillance was to improve the health and nutritional status at the district, particularly for children and women. Vital events such as births and deaths were recorded continuously over time, however, other events such as morbidity events were not. Like many other surveillance data, these were large in the number of subjects but without very detailed information on each subject. In the period between 1994 and 1998, there were about 15,000 households with around 8,000 children involved but the information on childhood morbidity was available for only a two week period every 90-th day.

Figure 1.1 is a typical event history collected in the surveillance. Data for certain events of interest (for instance, illness or death) are recorded for each child. The data is then available for investigating the determinants of childhood mortality and morbidity. Often in general surveillance data collection, observations can only be made partially because of technical or logistical reasons. Referring to Figure 1.1, as the surveillance was only conducted every 90-th day, the observation can only be recorded during period 1 and period 2. This common nature of event history data, known as *censoring* and *truncation*, has to be considered in the analysis.

Many specific epidemiological studies and trials are also conducted and organized under the surveillance system. One of them was the ZINAK study on zinc and iron supplementation in infants (Lind, 2004). This study was a community based, randomized, double-blind, placebo-controlled trial with the purpose to investigate the effect of four supplementation groups of iron, zinc, iron+zinc and placebo on iron, zinc status, infant growth, cognitive development and incidence of infant infectious diseases during the first six to twelve months of age. This thesis utilized the data on infant growth,

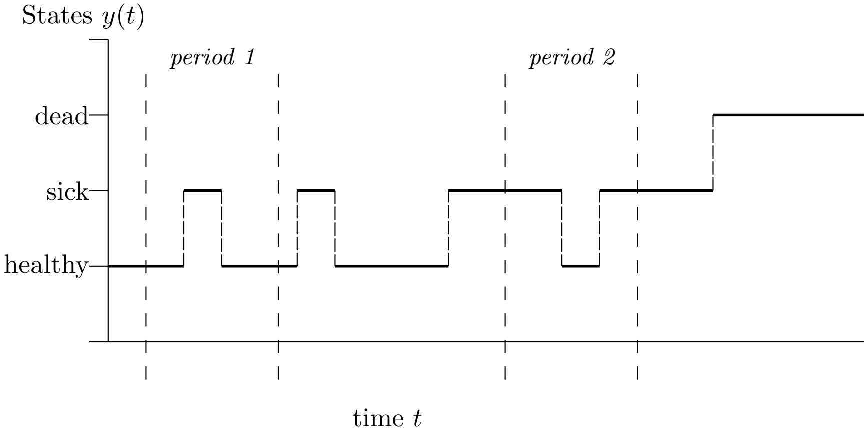


Figure 1.1: History of a hypothetical child experiencing healthy, ill and dead state, observed at two periods.

weight and infectious disease, respiratory infection. There were 680 infants aged six to twelve months participating in the study with daily supplementation and daily morbidity records, and monthly infant growth records.

Figure 1.2 shows an example of longitudinal data, repeated measurements of the weight of four infants across age in the ZINAK study. Here, the measurements are intermittently performed, once every month. One objective of the analysis of the data is to investigate the effect of the supplementations on weight development, taking into account other explanatory variables.

The study also considered morbidity (illness), such as respiratory infections. Figure 1.3 presents longitudinal measurements of weight together with the occurrence of respiratory infections. Interesting analyses of the data include studying the effect of supplementations on weight development, taking into account the respiratory infections as mentioned in the previous paragraph, or the effect of supplementa-

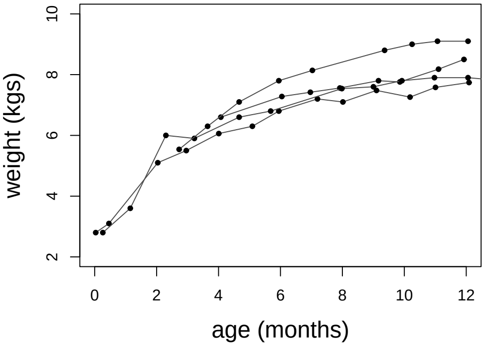


Figure 1.2: Repeated measurements on weight.

tions on the incidence of respiratory infections, taking into account weight development. The third possible analysis is to investigate weight and respiratory infection simultaneously, as both outcomes may actually affect each other, given the supplementations.

1.2 Review of the problem

Time-to-event analysis deals with the analysis of time measured from a well defined *time origin* up to the occurrence of a certain event of interest. The scale for measuring time can be ordinary clock time (minutes, days, years, and so forth) or other measurements such as *mileage* or *usage* which are common in reliability; *experience* or *exposure* which are common in epidemiology or social sciences.

Regression modeling of time-to-event data is commonly applied in studying the relationship between the outcome and independent (predictor) variables. The analysis can be performed through the density function or through the hazard function. As with many

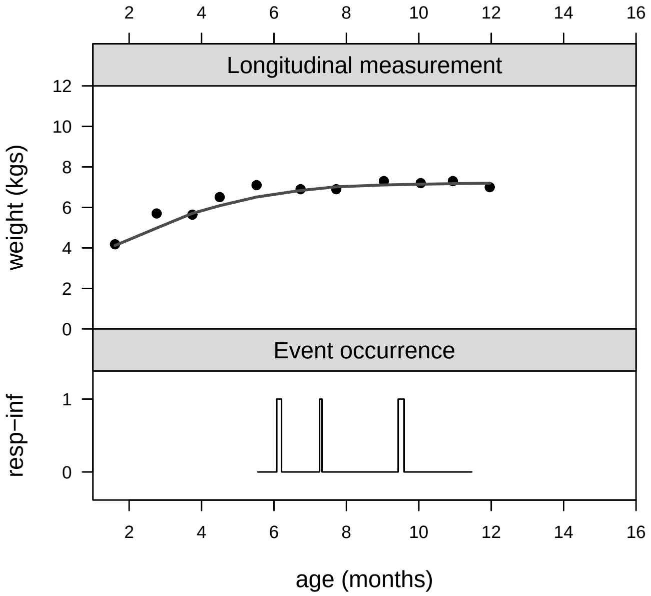


Figure 1.3: Repeated measurements on weight and respiratory infections for one infant.

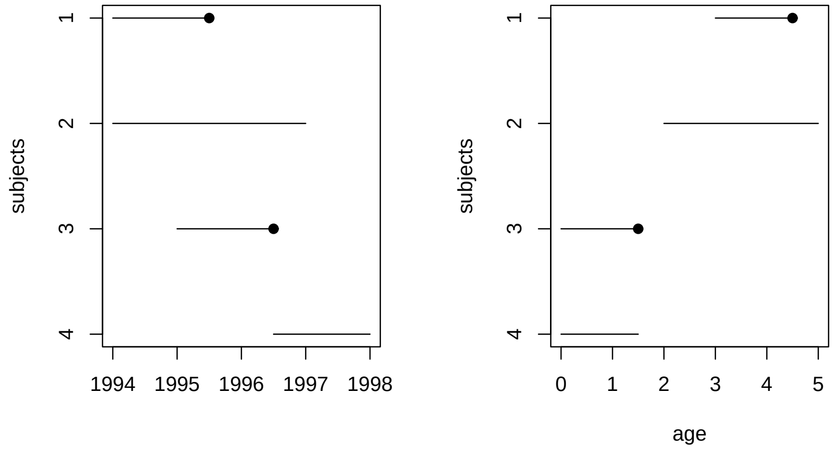
other statistical procedures, the analysis can be performed parametrically by specifying the density function, or non-parametrically by specifying nothing about the density function.

In this thesis, emphasis is given to the modeling of hazard functions using Cox's semiparametric model (Cox, 1972; Cox, 1975). The reasons of modeling the hazards are (Cox and Oakes, 1984; Hosmer and Lemeshow, 1999): (i) considering the immediate risk may be useful; (ii) comparisons of groups of individuals are sometimes sharpened by the hazard. For example, specific questions such as how survival is related to the treatments under study can be investigated by studying the estimated regression parameters from the hazard model; (iii) the hazard-based models can be extended to a more general event process, such as multiple events.

The semiparametric model is appealing in fields like epidemiology since most of the phenomena in epidemiological data are 'irregular' in the sense that a specific distribution function may not be easily determined. Furthermore, the idea of hazard comparison in the Cox model is similar to the well known *relative risk* in the common epidemiological analysis.

It has already been mentioned in the previous section that *censoring* and *truncation* are quite natural in event history data. Figure 1.4 gives a common description of censoring and truncation. The examples refer to the CHN-RL surveillance mortality data, for the period of time from 1994 to 1998, and for children under 5 years of age.

On the calendar time scale, many of the children did not enter the study at the beginning of the period in 1994 (subjects number 3 and 4, Figure 1.4(a). There is a similar situation in the age time scale where many of the children did not enter the study on their day of birth (subjects 1 and 2, Figure 1.4(b). This kind of missing information where the subjects are observed after the time origin



(a) Calendar time scale

(b) Age time scale

Figure 1.4: Four subjects with staggered entry (left-truncation), right-censored (the lines without dots) and event (the lines with dots) on two different time scales.

in the time-to-event data is known as *staggered entry*, *late entry* or *left-truncation*.

Some of the children experienced the events (deaths) and some of them were only partially observed, known as *censored*, because of the time limitation (only up to 1998 or reaching 5 years of age), and also due to other causes such as emigration.

Truncation and censoring may introduce several problems in the analysis such as *length biased sampling* (higher chance of being sampled for the longer survivors) and wasting information (if the analysis only utilize complete observations). Nowadays, time-to-event analysis can deal with these problem easily, for instance by using a counting process approach (Andersen, Borgan, Gill and Keiding, 1993; Therneau and Grambsch, 2000). The tools that facilitate truncation and censoring have made event history analysis with various time scales easier. For instance, the four subjects can be analyzed using a calendar time scale as easy as using an age time scale by specifying a counting process style of input (Therneau and Grambsch, 2000) corresponding to the scale used in the analysis. However, another complication may arise as discussed later.

Figure 1.5 represents the life experiences of the 4 subjects in Figure 1.4(a) and 1.4(b) on a *Lexis* diagram (Keiding, 1990). A Lexis diagram is a dual time scale system (usually calendar time and age), representing individual lives by line segments of unit slope, with events usually marked by dots. Representing the life experiences of the subjects in the period from 1994 to 1998 and under 5 years of age is clearer in this diagram than in the separate time scales of Figure 1.4(a) and 1.4(b).

Event history analysis often involves data with more than one time scale as shown in Figure 1.5. One early paper discussing this problem gave an example on the choice of time scale between age and age at first child's birth of women with breast cancer (Farewell and Cox, 1979). Another famous example is the *age-period-cohort*

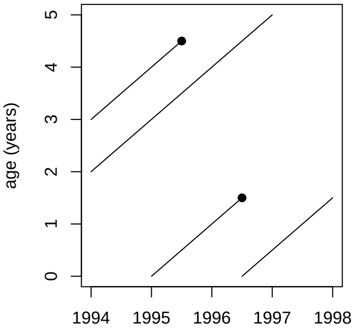


Figure 1.5: Four subjects with staggered entry (left-truncation), right-censored (the lines without dots) and event (the lines with dots) on a Lexis diagram.

model (Holford, 1998) which is popular in demography but carries an identification problem.

The multiple time scales problem also arises in *multi-state models* when many time scales are involved in the transition between states. Coping with several time scales is one of the challenges of multi-state models in epidemiology (Commenges, 1999).

Multiple time origins may be a more appropriate term than multiple time scales, since this problem deals with life experiences measured from many different origins (birthdate, starting date of surveillance, etc.). However, many authors have used the term multiple time scales in reference to this problem (Farewell and Cox, 1979; Berzuini and Clayton, 1994a; Oakes, 1995; Duchesne, 1999; Efron, 2002) and we continue to use the term.

This thesis considers the multiple time scales problem in the event history analysis as the first problem. This first problem includes the procedure to choose the most relevant time scale and to simultaneously model time scales.

Typically, event history data, such as the ZINAK study mentioned in the previous section, will also include longitudinal measurements collected intermittently across time. For instance, the growth or nutritional status, such as weight, were measured among children together with the morbidity outcomes, such as respiratory infections. The second problem considered in this thesis is the dual outcomes of event occurrence and longitudinal measurement.

When weight is considered as the primary outcome, weight will be the response variable with the occurrence or the symptom duration of respiratory infections as an explanatory variable, possibly with some other variables. The analysis can then be done using the longitudinal analysis methods proposed by Diggle, Heagerty, Liang and Zeger (2002).

Complications may arise when respiratory infection is the outcome of interest and weight is to be included as one explanatory variable. In many applications, continuous measurements of a longitudinal covariate, such as weight in the ZINAK study, are usually only available at some finite number of measurement times. This, potentially, becomes a problem in the ordinary Cox regression, since the method requires all values of covariates to be available at event times. Compromising the analysis by using cases with complete values of covariates is possible, but will lead to bias in the estimated regression coefficient.

Several methods have been proposed to cope with the above problem. They are the last value carried forward (LVCF), elapsed time (TEL) (Bruijne, Cessie, Kluin-Nelemans and Houwelingen, 2001), two-stage (Tsiatis, DeGruttola and Wulfsohn, 1995) and joint model method (Wulfsohn and Tsiatis, 1997; Henderson, Diggle and Dobson, 2000; Tsiatis and Davidian, 2004). Some comparisons have been made for some methods. The most recent, and perhaps, comprehensive one is the investigation by Andersen and Liestøl (2003). No attempt, however, has been made to compare the methods for repeated events such as respiratory infection in the ZINAK study.

1.3 Objectives and scope

The focus of this thesis is on the analysis of event history data using Cox's proportional hazards model with the objectives

- to demonstrate the use of event history analysis in the analysis of infant and child mortality, morbidity and growth and to identify the methodological problems in the analysis,
- to propose procedures to choose a basic time scale,
- to discuss the connections between the methods for modeling dual time scales and to perform quantitative comparisons between them,
- to compare existing methods to deal with longitudinal measurements in the Cox model with two proposed methods.

1.4 Outline and summary

Chapter 2 provides technical reviews of event history and longitudinal analysis. The concept of time-dependent covariates, which plays an important role in this thesis, is reviewed more comprehensively than the other topics. Chapter 3 presents the application of event history and longitudinal data analysis to childhood mortality and morbidity data from the CHN-RL surveillance data, and application on respiratory infection and weight data from the ZINAK study. This chapter gives the background to problems considered in the later chapters. Chapter 4 is devoted to the problem of multiple time scales. The procedures to choose the most relevant time scale and to model dual time scales are discussed. Simulation studies and application to infant mortality data are provided. Chapter 5 presents comparison of the methods to deal with longitudinal measurements in the event history analysis. An application to the infant respiratory infection and weight data is provided. Chapter 6 summarize and concludes this thesis and features further research and work in this area.

Chapter 2

Basic Methods

2.1 Introduction

This chapter is a brief technical exposition of basic theories and methods used for further developments in the later chapters. Longitudinal data analysis (LDA) and event history analysis (EHA) have similarities; for instance, in the nature of the data involved as mentioned in the previous chapter. The methods have many overlapping techniques and areas (see, for example, the review paper by Doksum and Gasko (1990), among others). The classical books on survival analysis and counting process theory by Cox and Oakes (1984); Kalbfleisch and Prentice (2002); Andersen et al. (1993) and the book on LDA by Diggle et al. (2002) are the main references for this chapter. This chapter also presents the similarities between the two analyses, especially for topics related to the time dependent covariates.

2.2 Event history analysis

2.2.1 Hazard and survival

Generic survival data is in the form of (T, δ), where $T = \min(T_e, T_c)$, the minimum of time to event T_e (such as failure or death time) and time to censored T_c; $\delta = I_{\{T_e \leq T_c\}}$, the indicator has a value of 1 if the event is observed or 0 if it is censored. Most often, we are also interested in including covariates in the data. The survival data becomes $(T, \delta, \mathbf{Z})$, where $\mathbf{Z} = (Z_1, \ldots, Z_p)'$ is a p-dimensional vector of covariates.

T is a non-negative random variable that can be continuous or discrete. We first consider the continuous case. There are many functions that describe the distribution of T. The cumulative distribution function $F(t) = \mathrm{P}(T \leq t)$ and the density function $f(t) = dF(t)/dt$ are the usual functions characterizing a random variable. More useful functions in survival analysis are the survivor function

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= \mathrm{P}(T \geq t), \end{aligned} \tag{2.1}$$

i.e., the probability of the duration time (e.g., lifetime) being longer than t, and the hazard function

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(t \leq T < t + \Delta t \mid T \geq t), \tag{2.2}$$

i.e., the probability of getting an event (e.g., death) within a short interval, conditional upon survival to time t.

Applying the definition of conditional probability and the relations between $F(t)$, $f(t)$, and $S(t)$, the relation between $\lambda(t)$ and

$S(t)$ can be derived as

$$\begin{aligned} \lambda(t) &= \frac{dF(t)}{dt}\frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

It also follows that

$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

and

$$S(t) = \exp\{-\Lambda(t)\}, \tag{2.3}$$

where

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.4}$$

is the integrated or cumulative hazard function.

As noted by Flemming and Lin (2000), observing (T, δ) rather than T_e give the crude hazard (Equation (2.2)) rather than the net hazard $\lambda_{\text{net}}(t) = \lim_{\Delta t \downarrow 0} \mathrm{P}(t \leq T < t + \Delta t \mid T_e \geq t)/\Delta t$. Therefore, in survival analysis the equality of the crude hazard and the net hazard is an important assumption. A sufficient condition for this assumption to be true is the independence of T_e and T_c.

2.2.2 The counting process approach

Aalen (1978) introduced a martingale-based approach to survival analysis, unifying the previously proposed non-parametric methods under a counting process framework. In this approach, survival data for a single subject i, (T_i, δ_i), is represented as $(N_i(t), Y_i(t))$, $t > 0$, where $N_i(t) = I_{\{T_i \leq t, \delta_i = 1\}}$ is the number of observed events in $[0, t]$ for subject i, and $Y_i(t) = I_{\{T_i \geq t\}}$ is the at-risk process.

The estimator of the cumulative hazard is based on the aggregated process $\widetilde{N}(t) = \sum N_i(t)$, the total number of events up to and including t and $R(t) = \sum Y_i(t)$, the *risk size* at time t. The estimator of the cumulative hazard (Equation (2.4)) is the Nelson-Aalen estimator, defined as

$$\hat{\Lambda}(t) = \int_0^t \frac{I_{\{R(u)>0\}}}{R(u)} \, d\widetilde{N}(u), \tag{2.5}$$

which intuitively can be thought of as the sum of the conditional probabilities that an event happens in the short intervals over $(0, t]$. The $d\widetilde{N}(t)$ can be decomposed as the discrete and continuous part $d\widetilde{N}(t) = \Delta\widetilde{N}(t) + n(t)dt$, where $d\Delta\widetilde{N}(t) = \widetilde{N}(t) - \widetilde{N}(t-)$ is the number of events occurring precisely at t for the discrete part and $n(t)$ is the change or differential for the continuous part.

An equivalent representation of the estimator is (Therneau and Grambsch, 2000)

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\Delta\widetilde{N}(t_i)}{R(t_i)}, \tag{2.6}$$

where $t_1, t_2, \ldots$ are the ordered event times.

The Nelson-Aalen estimator $\hat{\Lambda}(t)$ has a close connection to the Kaplan-Meier estimator (Kaplan and Meier, 1958). Let $\hat{S}(t) = \exp(-\Lambda(t))$ and $d\Lambda\hat{(}t_i) = d\widetilde{N}(t_i)/R(t_i)$, the increment in the Nelson-Aalen estimator at i-th event. Then when $\Delta\widetilde{N}(t_i)/R(t_i) \approx 0$,

$$\begin{aligned} \hat{S}(t) &= \prod_{i:t_i \leq t} \exp\{-d\hat{\Lambda}(t_i)\} \\ &\approx \prod_{i:t_i \leq t} \{1 - d\hat{\Lambda}(t_i)\}, \end{aligned}$$

which is the Kaplan-Meier product limit estimator.

Further, the process given by

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_i(u)du \tag{2.7}$$

is a *martingale* for subject i with respect to a proper filtration. (Aalen, 1978; Fleming and Harrington, 1991; Therneau and Grambsch, 2000) The martingale $M_i(t)$ (2.7) represents the difference between the observed and the model-predicted number of events over the interval $(0, t]$. Informally, a martingale with respect to a history $\mathcal{H}(t)$ is defined as a stochastic process that has a key property $E\{M(t) \mid \mathcal{H}(s)\} = M(s)$ for any $0 \leq s < t$.

We may rewrite (2.7) as $N_i(t) = \int_0^t Y_i(u)\lambda_i(u)du + M_i(t)$ and refer this decomposition as counting process=compensator+martingale, which is analogous to to data=model+noise in the statistical model decomposition (Therneau and Grambsch, 2000). This notion is important in studying residuals and diagnostics for survival models.

2.2.3 Regression models

Most often, it is desired to assess the effect of some covariates on survival. We need the time-to-event, event indicator and covariates information $(T, \delta, \mathbf{Z})$ for this analysis. The covariates may be fixed throughout the observation period (*time independent covariate*) or change with time (*time dependent covariate*).

The *Cox proportional hazards regression model* (Cox, 1972) is the most frequently used regression model in survival analysis. There are two approaches to this censored data regression model, the approach originally proposed by Cox and the counting process approach.

At this stage, we assume that the covariates are time independent. Let $S(t \mid \mathbf{Z})$ be the conditional survival function given the covariate vector $\mathbf{Z}$. The conditional hazard function is

$$\lambda(t \mid \mathbf{Z}) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(t \leq T < t + \Delta t \mid T \geq t, \mathbf{Z}). \tag{2.8}$$

When $\Delta t > 0$ is small, $\lambda(t \mid \mathbf{Z})\Delta t$ is approximately the conditional probability at event (failure, death) in the interval t to Δt given survival until time t and covariates $\mathbf{Z}$.

The Cox proportional hazards model specifies that

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{Z}), \tag{2.9}$$

where $\lambda_0(t)$ is an unspecified non-negative function called the *baseline hazard* common to all subjects, and $\boldsymbol{\beta}$ is a set of unknown regression coefficients.

Cox (1972; 1975) proposed a semiparametric approach for the proportional hazards model (2.9). Let $\mathcal{D}$ be the set of indices j of ordered event-times $t_1, t_2, \ldots, t_j, \ldots$ (For the moment we assume that only one subject gets an event at each event-time), and R_k be risk set at time t_k the subjects under observation and event-free immediately prior to t_k. The partial likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{k \in \mathcal{D}} \frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_k)}{\sum_{j \in R_k} \exp(\boldsymbol{\beta}'\mathbf{Z}_j)}, \tag{2.10}$$

in which the baseline hazard $\lambda_0(t)$ is canceled out. The $\boldsymbol{\beta}$ can be estimated using the *maximum partial likelihood*. Many researchers has investigated the large sample properties of this partial likelihood (see review by Fleming and Lin (2000)). If there is more than one event at a certain event-time (tied event-time), at least four procedures have been proposed to handle it (Therneau and Grambsch, 2000): *Breslow's approximation*, *Efron's approximation*, *exact partial likelihood*, and *averaged likelihood*. A method based on the maximum likelihood (ML) as an alternative of the maximum partial likelihood (MPL) is also proposed (Bailey, 1984; Broström, 2002). Efron's approximation is recommended since it is computationally feasible even with large tied data (Therneau and Grambsch, 2000). For heavier tied data, the ML estimator is superior (Broström, 2002).

The counting process approach treats the survival data in a more general way using the counting process notation $(N_i(t), Y_i(t))$ discussed earlier in this section. This generality is useful for a more elaborate survival analysis such as including time-dependent covariates, time-dependent strata, left truncation, multiple time scales, multiple events per subject, various problems with correlated data and case-cohort models. In the counting process approach, the partial likelihood is written as

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{n} \prod_{t>=0} \left[\frac{Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_k)}{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_j)} \right]^{dN_k(t)}, \tag{2.11}$$

where $Y_i(t)$ is zero-one at-risk process, and $dN_k(t) = 1$ if $N_k(t) - N_k(t-) = 1$, and $dN_k(t) = 0$ otherwise.

2.2.4 Diagnostics and stratification

As in ordinary linear regression, diagnostics are also important in the Cox regression model. There are a wide variety of model diagnostics available. Lindkvist (2000) has given an extensive review of the diagnostics and studied the added variable plot in the Cox model. For detecting the departure from the proportional hazards assumption, Schoenfeld residuals are useful (Grambsch and Therneau, 1994).

For certain situations, it is often necessary to stratify the subjects into disjoint groups when the proportionality assumptions do not hold for one or several covariates. In the stratified Cox model, the subjects in a certain stratum have a distinct baseline hazard function but common values for the regression coefficients. The partial likelihood for the stratified Cox model is given by

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{S} L_s(\boldsymbol{\beta}), \tag{2.12}$$

where S is the number of strata and $L_s(\boldsymbol{\beta})$ is the partial likelihood as in Equations (2.10) or (2.11) but calculated only for the subjects in stratum s.

2.2.5 Frailty

In a situation where the assumptions of independence and homogeneity of all individuals are violated, introducing *frailty models* may be useful (Andersen, 1991; Hougaard, 1995). Vaupel, Manton and Stallard (1979) introduced the term *frailty* in survival analysis. In the frailty model, an additional term is added to the Cox model of (2.9),

$$\lambda(t \mid W, \mathbf{Z}) = W\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}), \tag{2.13}$$

where W is the frailty term or the *random effect* term that is assumed to operate multiplicatively on the baseline hazard. Dependence and heterogeneity among individuals is modeled via this term by assuming W to follow a certain distribution. Estimation of W can be done using *penalized partial likelihood*, *EM algorithm* or the *Bayesian Gibbs sampler* approach (Sastry, 1997; Therneau and Grambsch, 2000; Manda, 2001).

2.2.6 Multistate models

The concepts and methods in survival analysis extend naturally to models with more than two states. For instance, the subjects may move among healthy, diseased and death states over time.

A multistate model is a stochastic process $\{X(t), t \in \mathfrak{T}\}$, with $X(t) \in S$ and $\mathfrak{T} = [0, \tau)$, $\tau \leq +\infty$. $X(t)$ denotes the state occupied by a subject at time t and $S = \{0, 1, \ldots, m\}$ is a finite state space.

The process starts with the initial distribution $\pi_j(0) = \mathrm{P}(X(0) = j)$, $j \in S$. As the process develops, a *history* (also called a *filtration*) $\mathcal{H}(t)$ will be generated containing all information about the process over interval $[0, t)$, such as the number of transitions until t (a counting process).

The multistate process is governed either by the *transition prob-*

abilities from state j to state k, defined as

$$\mathrm{P}_{jk}(s,t) = \mathrm{P}(X(t) = k \mid X(s) = j, \mathcal{H}(s-)) \tag{2.14}$$

for $j, k \in S$, $s, t \in \mathfrak{T}$, $s \leq t$; or by the *transition intensities* given the history just before t, $\mathcal{H}(t-)$, defined as

$$\alpha_{jk}(t \mid \mathcal{H}(t-)) = \lim_{\Delta t \to 0} \frac{\mathrm{P}_{jk}(t, t + \Delta t)}{\Delta t}. \tag{2.15}$$

A state $j \in S$ is *absorbing* if for all $t \in \mathfrak{T}$, $k \in S$, $j \neq k$, $\alpha_{jk}(t) = 0$, otherwise j is *transient.*

Here of course, we will always assume that the limits in the definition of the transition intensities $\alpha_{jk}(t \mid \mathcal{H}(t-))$ exist. Another assumption that may be applied to $\alpha_{jk}(t \mid \mathcal{H}(t-))$ is the *non-homogeneous Markov* assumption, $\alpha_{jk}(t \mid \mathcal{H}(t-)) = \alpha_{jk}(t)$, ignoring the history but still depending on time. A stronger assumption is the *homogeneous Markov*, which ignores both time and history, $\alpha_{jk}(t) = \alpha_{jk}$. In certain applications, it is possible to assume that the transitions depend on the time spent in the states, which leads to the *semi-Markov* assumption.

2.3 Longitudinal data analysis

2.3.1 Notation and approaches

Longitudinal data sets consist of a measurement (outcome or response) variable Y_{ij} and vector of explanatory variables $\mathbf{x}_{ij}$ observed at time t_{ij} for subject $i = 1, \ldots, m$ and observation $j = 1, \ldots, n_i$. The mean and variance of Y_{ij} are denoted by $\mathrm{E}(Y_{ij}) = \mu_{ij}$ and $\mathrm{Var}(Y_{ij}) = v_{ij}$. For each subject i, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ denotes the vector of measurements with mean $\mathrm{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and $n_i \times n_i$ covariance matrix $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{V}_i$. The covariance between Y_{ij} and Y_{ik} is denoted by $\mathrm{Cov}(Y_{ij}, Y_{ik}) = v_{ijk}$. The $n_i \times n_i$ correlation matrix of $\mathbf{Y}_i$ is denoted by $\mathbf{R}_i$. The complete $N = \sum_{i=1}^{m} n_i$ measurements are denoted by $\mathbf{Y} = (Y_i', \ldots, Y_m')'$ with mean $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and variance matrix $\mathrm{Var}(\mathbf{Y}) = \mathbf{V}$.

The scientific question of interest could be the pattern of change over time of the outcome or the dependence of the outcome on the covariates. Most of the approaches of LDA consider regression models under general linear model or the extension of generalized linear model.

2.3.2 General linear models

We consider the data setup and notations as described in the previous section. Under the general linear model, it is assumed that $\mathbf{Y}$ has a multivariate Normal distribution

$$\mathbf{Y} \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{V}). \tag{2.16}$$

This longitudinal data model is completed by specifying the form of mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{V}$.

The mean $\boldsymbol{\mu}$ is specified as a linear model

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{2.17}$$

with $\mathbf{X} = (x_{ij1}, \ldots, x_{ijp})$ are $N \times p$ design matrix that may include covariate of interests and functions of time, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a p-vector of unknown regression coefficients.

The specification of $\mathbf{V}$ can be made to include at least three different sources of random variation: *random effects*, *serial correlations* and *measurement errors*. A model that incorporates all the three sources of variation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \mathbf{W}(\mathbf{t}) + \boldsymbol{\epsilon}, \tag{2.18}$$

where $\mathbf{U}$, $\mathbf{W}(t)$ and $\boldsymbol{\epsilon}$ correspond to random effects, serial correlations and measurement errors, respectively; $\mathbf{Z}$ is the design matrix of $\mathbf{U}$; $\mathbf{t} = \{t_{ij}\}$ is a set of times at which the measurements are made. Altogether, $\mathbf{U}$, $\mathbf{W}(t)$ and $\boldsymbol{\epsilon}$ has zero mean and specifies the variance matrix $\mathbf{V}$ of model (2.16).

To be precise, it is assumed that $\mathbf{U} \sim \text{MVN}(0, \boldsymbol{\Psi})$, $\boldsymbol{\epsilon} \sim N(0, \tau^2)$ and $\mathbf{W}(\mathbf{t})$ are independent stationary Gaussian processes with mean zero, variance σ^2 and correlation function $\rho(u)$ which still needs to be parameterized further. For instance, the popular choice of $\rho(u)$ is $\rho(u) = \exp(-\phi u^c)$ with $c = 1$ (the exponential correlation) or $c = 1$ (the Gaussian correlation) and $\phi > 0$ (Diggle, 1988).

For each individual i, the covariance matrix $\mathbf{V}_i$ can be written as

$$\mathbf{V}_i = \mathbf{Z}_i \boldsymbol{\Psi} \mathbf{Z}_i' + \sigma^2 \mathbf{H}_i + \tau^2 \mathbf{I}_i, \tag{2.19}$$

where $\mathbf{H}_i$ is the $n_i \times n_i$ symmetric matrix with the (j,k)-th element $h_{ijk} = \rho(\mid t_{ij} - t_{ik} \mid)$, and $\mathbf{I}$ is the $n_i \times n_i$ identity matrix.

The specification of $\mathbf{V}_i$ will lead to various linear models, from the simple classical linear model with independent errors to more complicated ones, such as linear model that includes all those three sources of errors.

Several estimation methods for this longitudinal model has been proposed for the special case of variance structure given by (2.19) or for the general case. Laird and Ware (1982); Diggle et al. (2002) suggested *maximum likelihood* (ML) and *restricted maximum likelihood* (REML) with the remark that REML is usually better than ML. Goldstein (1986; 1989) suggested *iterative generalized linear model* (IGLS) and *restricted* IGLS (RIGLS) for more general multilevel structure. Bates and Pinheiro (1998) proposed EM estimation followed by Newton-Rhapson or quasi-Newton optimization of the log-likelihood or the log-restricted-likelihood. Bayesian methods also have been suggested, for instance using Gibbs sampling (Zeger and

Karim, 1991). The multilevel mixed models as a general case for the longitudinal models with normal and non-normal responses are reviewed in Section 2.3.4.

2.3.3 Generalized estimating equations

For a more general longitudinal model with non-Gaussian outcome, an extension of the generalized linear model (GLM) was suggested by Liang and Zeger (1986). Like the ordinary GLM (McCullagh and Nelder, 1989), the model can handle a wide range of discrete and continuous outcome distributions such as binomial, Poisson, gamma and normal.

Using the notation and data setup introduced in Section 2.3.1, in this model the mean of $\mathbf{Y}_i$ is specified as

$$\boldsymbol{\mu}_i = h(\mathbf{X}_i\boldsymbol{\beta}), \tag{2.20}$$

where $\boldsymbol{\beta}$ is p-vector of unknown parameters. The inverse of h is known as the ”link” function in the GLM terminology. The variance of $\mathbf{Y}_i$ is specified through the $n_i \times n_i$ ”working” correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$. It is said to be ”working” since we do not expect it to be correctly specified (Zeger and Liang, 1986). The $\boldsymbol{\alpha}$ are some unknown parameters common to all subjects.

The working covariance matrix of $\mathbf{Y}$ is

$$\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}/\phi, \tag{2.21}$$

where $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with known function $g(\mu_{ij})$ as the j-th diagonal element and ϕ is a scale parameter.

The *generalized estimating equation* (GEE) of this longitudinal data model is given by

$$\sum_{i=1}^{m} \mathbf{D}_i' V_i^{-1}\mathbf{S}_i = \mathbf{0}, \tag{2.22}$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $\mathbf{S}_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$. The GEE estimator of $\boldsymbol{\beta}$ is the solution of equation (2.22). Liang and Zeger (1986) studied the consistency of the estimator and proposed an iterative procedure to estimate $\boldsymbol{\beta}$.

A problem that frequently arises in longitudinal data is *missing values*. The GEE estimation is still consistent even when $\mathbf{R}_i$ is miss-specified provided that the missing values are completely at random (Liang and Zeger, 1986; Diggle et al., 2002). When the missing values are not completely random, joint modeling of dropouts (missing values) and longitudinal measurements may be needed.

The approach considered here is called the *population averaged* (PA) models (Zeger, Liang and Albert, 1988) in which the aggregate response for the population is modeled. Another approach is the *subject specific* (SS) models in which heterogeneity in regression parameters is modeled. The next section considers the second approach.

2.3.4 Generalized linear mixed models

The models discussed in the previous two sections can be extended to more general class of models. Generalized linear mixed model (GLMM) is an extension of GLM by including random effects, or more general multilevel or hierarchical structure in the model.

Rather than modeling the mean of $\mathbf{Y}$ as in the previous section, this model focus on modeling $\mathbf{u}_i$ =E($\mathbf{Y} \mid \mathbf{b}$) specified as

$$\mathbf{u}_i = h(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i), \tag{2.23}$$

where $\mathbf{b}$ is vector of random effects with design matrix $\mathbf{Z}_i$. The inverse of h is the "link" function as in Equation (2.20). This model is also known as subject specific (SS) in (Zeger et al., 1988). SS models are desirable when the response of an individual is the focus rather than the average population response.

The GEE can be used for this model as well. In the GLMM both the link function and the random effects distribution must be correctly specified. To use GEE for the GLMM, the marginal moments $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$ of Equations (2.20) and (2.21) are calculated from the conditional moments and the random effects distribution F and solve the GEE.

The GLMM estimation using GEE aims primarily at estimating fixed effects and does not estimate the random component terms which are often useful for prediction or in model diagnostic. Lately, Lee and Nelder (2001) developed hierarchical GLM that allows models with any combination of GLM distribution for the response with any conjugate distribution for the random effects, structured dispersion components, different link functions for the fixed and random effects and the use of quasilikelihoods in place of likelihoods for either or both of the mean and dispersion models.

2.4 Time-dependent covariates

2.4.1 Some useful classifications

Longitudinal or event history data has the advantage of observing the temporal order of the outcome and covariate. The analysis of covariate changes may be useful in studying causal relationships. A time-dependent covariate is a covariate that vary over time. This section discusses basic issues of time-dependent covariates for both event history and longitudinal data.

In survival analysis, Kalbfleisch and Prentice (2002, Section 6.3) classify time-dependent covariates as *external* and *internal.* Let $\mathbf{x}_i(t)$ denote the time-dependent covariate at time t for individual i and $\mathbf{X}_i(t) = \{\mathbf{x}_i(u); 0 \leq u < t\}$ denote the covariate history up to time

t. For each individual i, the hazard function of (2.8) becomes

$$\lambda_i(t \mid \mathbf{X}_i(t)) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(t \leq T_i < t + \Delta t \mid T_i \geq t, \mathbf{X}_i(t)). \qquad (2.24)$$

An external (time-dependent) covariate $\mathbf{X}_i(t)$ satisfies the condition

$$\mathrm{P}(u \leq T_i < u + \Delta u \mid T_i \geq u, \mathbf{X}_i(u)) = \\ \mathrm{P}(u \leq T_i < u + \Delta u \mid T_i \geq u, \mathbf{X}_i(t)) \qquad (2.25)$$

for all u, t such that $0 < u \leq t$. An equivalent condition is

$$\mathrm{P}(\mathbf{X}_i(t) \mid T_i \geq u, \mathbf{X}_i(u)) = \mathrm{P}(\mathbf{X}_i(t) \mid T_i = u, \mathbf{X}_i(u)), \quad 0 < u \leq t. \qquad (2.26)$$

This condition implies that the future path of $\mathbf{X}_i(t)$ up to any time $t > u$ is not affected by the occurrence of an event at time u.

When the conditions (2.25) or (2.26) are not satisfied, $\mathbf{X}_i(t)$ is called an internal covariate. The main consequence of internal covariate is that the future path of the covariate is affected by the event occurrence.

External covariates may be classified further as *fixed*, *defined* and *ancillary* covariates. When the external covariate is fixed across time, e.g., $\mathbf{X}(t) = \mathbf{Z}$, then the hazard function of (2.24) is the same as (2.8). A defined covariate is when $\mathbf{X}(t)$ determined in advanced for each individual. This covariate is usually a factor determined in experimental study. Another example is the age of individual or calendar time across the study. An ancillary covariate is the output of stochastic processes that is external to the time-to-event process of the individual, such as pollution, seasonality or social-economics conditions.

The relation between the hazard function and the survival function for the external covariate is given by

$$S(t \mid \mathbf{X}(t)) = \exp\left[-\int_0^t \lambda(u \mid \mathbf{X}(u))du\right], \tag{2.27}$$

which is similar to that of a time-independent covariate. The relationship for the internal covariate is different to (2.27) and discussed in the next section.

In LDA, there are similar definitions for internal and external covariates. We consider the notation in Section 2.3.1 with modification, X_{ij} denotes the time-dependent covariate and $\mathbf{Z}_{ij}$ denotes the time-independent covariates. Here j represents discrete follow-up times. Adapted from econometrics terminology, in the LDA, a covariate is classified as *exogenous* or *endogenous* (Diggle et al., 2002).

Define the history of time-dependent covariates and outcomes for individual i up to time t as $\mathcal{H}_{Xi}(t) = \{X_{i1}, X_{i2}, \ldots, X_{it}\}$ and $\mathcal{H}_{Yi}(t) = \{Y_{i1}, Y_{i2}, \ldots, Y_{it}\}$, respectively, exogenous is defined as

$$f(X_{it} \mid \mathcal{H}_{Yi}(t), \mathcal{H}_{Xi}(t-1), \mathbf{Z}_i) = f(X_{it} \mid \mathcal{H}_{Xi}(t-1), \mathbf{Z}_i), \tag{2.28}$$

where $f(.)$ represents a density or probability function of the covariate. When the condition (2.28) is not satisfied, $\mathcal{H}_{Xi}(t)$ is endogenous.

When covariates are exogenous, the future of the covariates are not affected by the outcomes and the analysis can focus on specifying the dependence of Y_{it} on $X_{i(t-1)}, X_{i(t-2)}, \ldots$. Generally, the approach consider $\mathrm{E}(Y_{it} \mid X_{is}, s < t)$. For example, a GEE model with single lagged covariate can be specified as

$$h(E(Y_{it} \mid X_{is}, \mathbf{Z}_i)) = \beta_0 + \beta_1 X_{i(t-k)} + \boldsymbol{\beta}_2' \mathbf{Z}_i. \tag{2.29}$$

All methods and inferences discussed in Section 2.3.2 and Section 2.3.3 basically can be used in the lagged model.

2.4.2 Approaches in the Cox model

The partial likelihood for the Cox model with time-dependent covariate is similar with (2.11). The form of the Cox partial likelihood is

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{n} \prod_{t>=0} \left[\frac{Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_k(t))}{\sum_{j=1}^{n} Y_j(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_j(t))} \right]^{dN_k(t)}, \tag{2.30}$$

where $\mathbf{Z}_j(t)$ is the time-dependent covariate at time t. The calculation of the likelihood requires covariate values at the event times.

Typical situations in survival analysis with time dependent covariates are illustrated in Figure 2.1. Figure 2.1(c) is a *switching treatments* time dependent covariate (Cox and Oakes, 1984, Chapter 8) in which subjects may change from one treatment to another. The usual method to deal with such a covariate, given that the covariate is external, is to split the individual life time by the time when the covariate values change. This is easy to manage in standard statistical packages that facilitate the counting process style of input.

Figure 2.1(b) is an example of a *defined* time-dependent covariate. For example, if the time scale used in the analysis is time since entering the study, a defined covariate could be the age of the individuals. Of course, age has the same speed as the survival time, and their values are always available at any event time. Unlike the previous example, it is computationally more efficient to split the individual life times by event times.

Often, covariates are collected intermittently across the time such that their values are not available at the event times (Figure 2.1(a)). In this situation several methods have been proposed. These include the *last value carried forward* (*LVCF*) method, using the last value of the covariate to substitute the missing value prior to the event time.

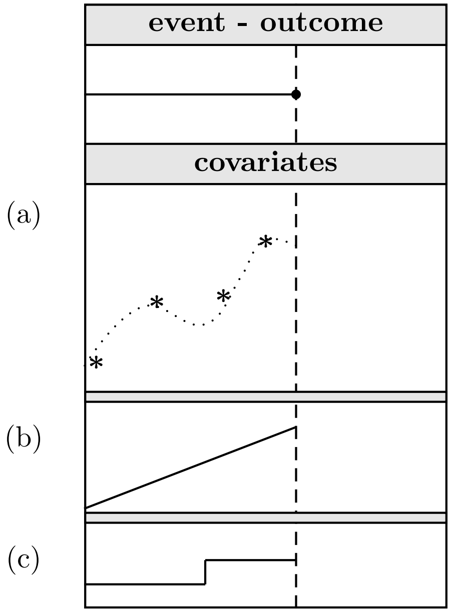


Figure 2.1: Time-to-event and time-dependent covariates: (a) intermittently observed (b) *defined* covariate (c) *switching treatments* covariate.

Imputation methods such as *two-stage* estimation and smoothing can be applied to this problem as well. In the two-stage method, a mixed model is fitted to the data at each event time with time-dependent covariate as the response (Pawitan and Self, 1993; Tsiatis et al., 1995). Bruijne et al. (2001) suggested another approach using *time elapsed since the last measurement* (TEL) in the Cox's regression model together with the LVCF or other methods of imputation. The TEL can be considered as "the age of the longitudinal measurement" in which Cox's model that includes TEL may be better than the Cox's model with only LVCF or two-stage imputation.

More general methods based on the joint modeling of event-times and longitudinal measurements have also been proposed (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Lin, Turnbull, McCulloch and Slate, 2002; Xu and Zeger, 2001; Tsiatis and Davidian, 2004). Basically, this model consider two linked sub-models, one for the longitudinal measurements model and one for the event-time model. The two sub-models are joined together with a Gaussian latent process. Without the latent process the models become the ordinary separate longitudinal measurement and event-time models.

To estimate the model, a likelihood based method leading to EM algorithms has been proposed (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Lin, Turnbull, McCulloch and Slate, 2002). Other methods are based on a Bayesian approach (Faucett and Thomas, 1996; Xu and Zeger, 2001; Guo and Carlin, 2004). Utilizing the usual connection between survival analysis and GLM, the model can also be estimated using the GEE approach (Rochon and Gillespie, 2001) and by generalized linear latent mixed models (Rabe-Hesketh, Yang and Pickles, 2001).

2.4.3 Time-dependent confounders

The notion of time-dependent confounders in epidemiology has been recognized at least by Robins (1986) and later in the epidemiological journals in the 90's (see for example articles by The Cebu Study Team (1991); Pearce (1992); and Zohoori and Savitz (1997)). Keiding (1999) gave an overview of this problem in event history analysis. A time-dependent confounder, often arising in longitudinal or cohort studies, is both a confounder and an intermediate variable. It is also known as feedback models (Zeger and Liang, 1991) and related to the internal or endogenous discussed covariates in the previous section.

To deal with time-dependent confounders in longitudinal data, we may use a method proposed by Zeger and Liang (1991). The method is based on GEE models allowing for both lagged response and endogenous covariates. A more general solution with theoretical exposition can be found in a book by van der Laan and Robins (2003).

For EHA, time-dependent confounders is closely related to internal covariates. The hazard function for an internal covariate is defined by (2.24) but conditioned on the time-dependent covariate only up to $t-$ (time just before t) and not further. The relation (2.27) does not hold. In fact, for survival data, the internal covariate requires the survival of individuals for its existence, therefore the survival function is always one, provided that $\mathbf{x}(t-) \neq 0$. Generally the survival function will be (Jewell and Kalbfleisch, 1996; Andersen, 2003)

$$S(t \mid \mathbf{X}(t)) = \mathrm{E}\left\{\exp\left[-\int_0^t \lambda(u \mid \mathbf{X}(u))du\right]\right\}, \qquad (2.31)$$

where the expectation is taken with respect to the sample path $\mathbf{X}(.)$. The marginal survival probability at t given the past history is the average over the possible paths among individuals at risk for $\mathbf{X}(t)$.

In Cox's regression model, care must be taken in interpreting the estimated coefficients, since $\mathbf{X}(t)$ may serve as an intermediate variable. However, an internal covariate is not something to be avoided, a particular kind of internal covariates known as *marker* or *surrogate end-point* have many useful applications (Jewell and Kalbfleisch, 1996; Prentice, 1989).

The multiple time scales problem in the next chapter is closely related to the *defined* covariate (Figure 2.1(b)), whereas the longitudinal measurement problem in Chapter 5 is closely related to the *intermittently observed* time-dependent covariate (Figure 2.1(a)).

Chapter 3

Analysis of Childhood Mortality, Morbidity and Growth

3.1 Introduction

This chapter presents some applications of event history analysis (EHA) and longitudinal data analysis (LDA) to a childhood epidemiological study. The Community and Health Nutrition Laboratories (CHN-RL) surveillance and the ZINAK study on zinc and iron supplementation in infants introduced in Chapter 1 are the two main sources of data used in the analysis. This chapter is also meant to be a natural background for methodological development in the later chapters.

3.2 Mortality

Child survival in developing countries has been investigated intensively, especially since the study by Mosley and Chen (1984). The Cox model for analyzing childhood mortality in developing countries has been employed by, among others, Trussell and Hammerslough (1983) and Pebley and Stupp (1987). Using the Community Health and Nutrition Research Laboratories (CHN-RL) data, infant mortality has been investigated relating to the effects of sibling status (Wahab, Winkvist, Stenlund and Wilopo, 2001). In general, they concluded that boys had higher infant mortality rates than girls although the difference was not great. The risk for boys was even higher when they were born after a few siblings compared with being first-born. Further study is still needed to evaluate the different mortality pattern among boys and girls in that area.

Here, we investigated more aspects on the effect of siblings and gender on childhood mortality, taking into account clustering levels of mother, household, community and village using EHA. Detail of the analysis has been reported elsewhere by Danardono (2003).

3.2.1 Data, study variables and models

Rather than considering the live births for a period of 1995 to 1996 in the CHN-RL surveillance (Wahab et al., 2001) as the subjects, we considered all children observed since the start of surveillance on October 1994. This scheme has an advantage in utilizing all information available in the surveillance but introduces length-biased sampling (Section 1.2). Consequently, the length-biased sample selection has to be taken into account in the analysis by using left-truncation. After excluding some twins and incomplete records, 7889 children were available in the data set with 2948 of them being born after the start of the surveillance data collection.

Specifically, we investigated the *sibling* and *gender* effects on mortality. The sibling factor has been pointed out as being of interest, in the way that it may explain the difference in care between boys and girls and possible competing resources among them (Wahab et al., 2001). To study this effect, several variables were constructed based on *gender* and *birth order*. The sibling variable is a time-dependent covariate, a ”switching treatment” like covariate (see Figure 2.1(c) in Chapter 2).

We give one example of this variable construction. We use the term *index child* to denote the child under consideration. Suppose we have information as in Figure 3.1(a). When a younger sibling was born the value of this time dependent covariate is changed from 0 to 1. We may further consider the gender of the younger sibling and categorize *boy* or *girl* rather than just 1 as the value of this time-dependent covariate.

In Figure 3.1(a), there are two children who experienced the events before the event times of the index child, and one child, the sibling of the index child, who has not experienced the event. We can construct the data suitable for event history analysis using Cox's model by event-time splitting (Figure 3.1(b)) or covariate-time splitting (Figure 3.1(c)). Both constructions will lead to the same result. However, in the case of switching treatment covariate, in which the value of the covariate is a step function with only a few values, splitting by covariate times is more efficient since it usually gives less splitting intervals than event-time splitting.

Another situation is when the index child did not enter from birth (delayed entry or left-truncation) and the younger sibling was born before the entry time. In this case, there is no splitting by the younger sibling covariate, except if the sibling dies. A similar construction is applied for the older sibling covariate where the value is changed when the older sibling dies. For this analysis, we only constructed covariates for the closest sibling (one younger or one older sibling).

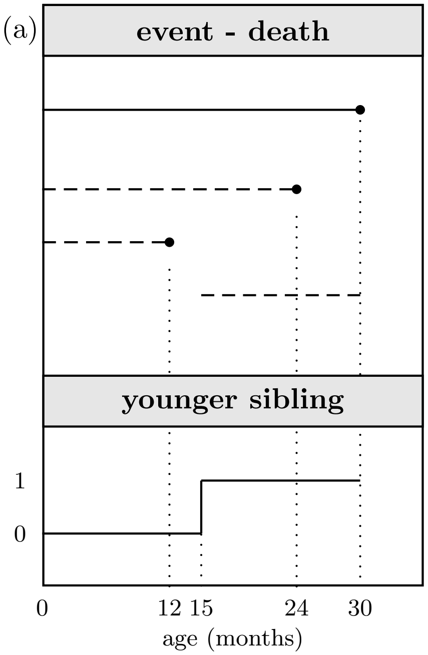
We used the Cox proportional hazards model reviewed in Section 2.2.3, i.e., the standard model of Equation (2.9) and the shared frailty model of Equation (2.13). We used gamma frailty to model the frailties. Currently, there is no general agreement about the best frailty distribution for practical frailty modeling (Therneau and Grambsch, 2000). The Gamma distribution, however, has been used in several statistical and demographical studies (Guo and Rodríguez, 1992; Sastry, 1997). To estimate the frailty term, we used the penalized partial likelihood approach (Therneau and Grambsch, 2000), available in the ***R*** survival package (Ihaka and Gentleman, 1996; R Development Core Team, 2004).

3.2.2 Results

We obtained two hazard models for the childhood mortality: the *infant mortality* (0-1 year of age) and *child mortality* (1-5 years of age), presented in Table 3.1 and 3.2, respectively.

For the infant mortality hazard model, the strongest, yet unsurprising, result is the effect of *maternal education*. Higher education gave a protective effect for childhood mortality. The *gender* of the index child alone was slightly a significant factor for childhood mortality; girls seemed to have lower risk than boys. *Birth order* also shows a significant linear effect on mortality, the risk increases with higher birth order. The *older sibling* variable does not seem show any effect, the relative risk of infants (0-1 year of age) who had no older sibling, older brother or sister are the same.

After infancy (aged 1-5 years), the effects of *gender*, *birth order* and *maternal education* seem to disappear, on the other hand the effects of siblings appear. We also examined the interaction between gender of the index child and the gender of the older sibling as well



(b)

start	stop	status	sibling
0	12	0	0
12	24	0	1
24	30	1	1

(c)

start	stop	status	sibling
0	15	0	0
15	30	1	1

Figure 3.1: Sibling as a time-dependent covariate: (a) The bold line under event-death frame is the *index child*, the dashed lines are other children; the line under younger sibling frame is the time-depedent covariate value; (b) splitting by event times; (c) splitting by covariate times.

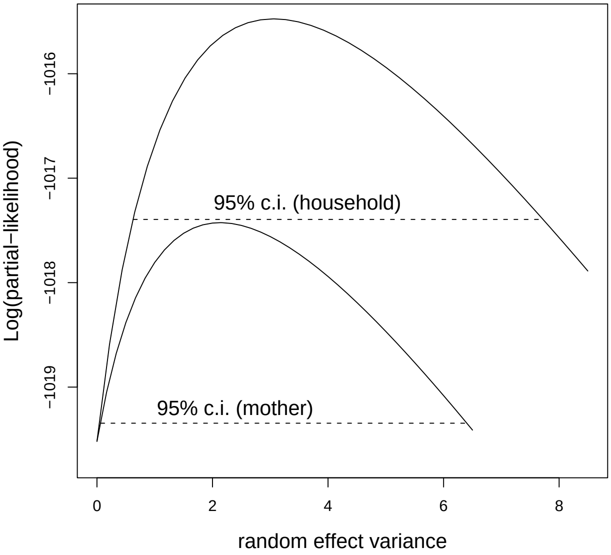


Figure 3.2: Profile likelihood for the *mother* and *household* random effect variance for infant mortality model.

as the younger sibling. Neither interaction was significant. The risk of mortality is higher when the index child (boy or girl) has an older or younger brother. The above results probably do not reflect gender difference in care, in favor of boys, since the index child with the higher risk is either boy or girl, but it may reflect exhausting resources when a family has a boy (or boys) that lead to childhood mortality. The confidence intervals of the relative risks of this model are shown in Table 3.2, under the standard model. The estimates are rather poor with wide confidence intervals for the sibling variable and maternal education.

We also included several frailty terms that assumed to operate on a certain meaningful level. The mother frailty may capture any unobserved variables that operate on children born from the same

Table 3.1: Five hazard models for infant mortality (0-1 years)

Variables	standard model	mother frailty	household frailty	community frailty	village frailty
	RR (c.i.)	RR (c.i.)	RR (c.i.)	RR (c.i.)	RR (c.i.)
Gender					
boy	1	1	1	1	1
girl	0.71 (0.50-1.01)	0.70 (0.49-0.99)	0.69 (0.49-0.98)	0.71 (0.50-1.01)	0.72 (0.51-1.01)
Birth order (linear)	1.16 (1.00-1.35)	1.17 (1.01-1.36)	1.18 (1.01-1.37)	1.16 (1.00-1.34)	1.16 (1.00-1.34)
Maternal age at delivery					
20-29 year	1	1	1	1	1
$<$ 20 year	1.72 (0.89-3.32)	1.76 (0.91-3.40)	1.79 (0.93-3.45)	1.69 (0.88-3.26)	1.70 (0.88-3.28)
+ 30 year	0.95 (0.61-1.48)	0.98 (0.63-1.53)	0.99 (0.64-1.54)	0.97 (0.62-1.50)	0.96 (0.62-1.48)
Older sibling					
none	1	1	1	1	1
older brother	1.29 (0.72-2.34)	1.23 (0.68-2.22)	1.20 (0.67-2.17)	1.29 (0.71-2.32)	1.29 (0.72-2.34)
older sister	1.07 (0.57-1.99)	1.02 (0.55-1.90)	0.99 (0.54-1.87)	1.06 (0.57-1.98)	1.07 (0.57-1.99)
Maternal education					
12 years of education	1	1	1	1	1
no education	11.59 (3.84-34.96)	11.83 (3.92-35.69)	11.98 (3.97-36.14)	11.06 (3.67-33.37)	11.47(3.80-34.60)
6 years of education	5.95 (2.17-16.33)	6.07 (2.21-16.64)	6.08 (2.22-16.69)	5.8 (2.11-15.89)	5.96(2.17-16.34)
9 years of education	4.63 (1.56-13.72)	4.73 (1.59-14.02)	4.76 (1.61-14.11)	4.59 (1.55-13.61)	4.63(1.56-13.74)
Variance of random effect[1]					
mother		2.135 (0.041)			
household			3.074 (0.004)		
community				0.103 (0.319)	
village					0.054 (0.334)

[1] The estimated variance of random effects and the *p*-value of the LRT

Table 3.2: Five hazard models for child mortality (1-5 years)

Variables	standard model	mother frailty	household frailty	community frailty	village frailty
	RR (c.i.)	RR (c.i.)	RR (c.i.)	RR (c.i.)	RR (c.i.)
Gender					
boy	1	1	1	1	1
girl	0.97 (0.46-2.07)	0.97 (0.46-2.07)	0.97 (0.46-2.07)	0.97 (0.46-2.07)	0.96 (0.45-2.04)
Birth order (linear)	0.87 (0.58-1.31)	0.87 (0.58-1.31)	0.87 (0.58-1.31)	0.87 (0.58-1.31)	0.87 (0.58-1.30)
Maternal age at delivery					
20-29 year	1	1	1	1	1
$<$ 20 year	1.31 (0.19-8.90)	1.31 (0.14-12.10)	1.31 (0.14-12.1)	1.31 (0.14-12.10)	1.33(0.14-12.28)
$+$ 30 year	0.84 (0.38-1.88)	0.84 (0.34-2.06)	0.84 (0.34-2.06)	0.84 (0.34-2.06)	0.85 (0.35-2.09)
Older sibling					
none	1	1	1	1	1
older brother	11.46 (2.64-49.8)	11.46 (2.01-65.23)	11.46 (2.01-65.23)	11.46 (2.01-65.23)	11.57(2.03-65.82)
older sister	6.34 (1.26-32.01)	6.34 (1.03-38.87)	6.34 (1.03-38.87)	6.34 (1.03-38.87)	6.35(1.04-38.97)
Younger sibling					
none	1	1	1	1	1
younger brother	4.86 (1.44-16.45)	4.86 (1.45-16.31)	4.86 (1.45-16.31)	4.86 (1.45-16.31)	4.88(1.46-16.38)
younger sister	1.2 (0.16-9.01)	1.2 (0.15-9.65)	1.2 (0.15-9.65)	1.2 (0.15-9.65)	1.17 (0.14-9.39)
Maternal education					
12 years of education	1	1	1	1	1
no education	2.03 (0.12-33.78)	2.03 (0.13-32.95)	2.03 (0.13-32.95)	2.03 (0.13-32.93)	1.93(0.12-31.33)
6 years of education	5.02 (0.68-37.29)	5.02 (0.67-37.67)	5.02 (0.67-37.67)	5.02 (0.67-37.67)	4.95(0.66-37.12)
9 years of education	2.07 (0.19-22.09)	2.07 (0.19-22.92)	2.07 (0.19-22.92)	2.07 (0.19-22.92)	2.07(0.19-23.00)
Variance of random effect[1]					
mother		≈ 0 ≈ 1			
household			≈ 0 ≈ 1		
community				0.005 0.947	
village					0.184 0.423

[1] The estimated variance of random effects and the p-value of the LRT

mother, such as genetic factors and maternal competence. At the household level, family size, socio-economic status and housing condition may be captured by household frailty term. At the broader coverage of level, community and village level were also included. These terms will account for the possible effects of infrastructure, climate, and other environmental factors within the community; and institutional effect within the village.

Figure 3.2 shows the profile likelihood for the mother and household frailty term. The 95% confidence interval is constructed by referencing a horizontal line 3.84/2 units below the maximum log-partial likelihood. The reference line is obtained by assuming that 2×(the difference in likelihood) has Chi-square distribution with one degree of freedom. The maximum log likelihood of the household frailty model is -1015.48, which corresponds to the value 3.074 of the estimated random effect variance, and -1017.43 for the mother frailty effect, which corresponds to the estimated random effect variance of 2.14. The intervals range from 0.63 to 7.70 for the household frailty, and 0.07 to 6.38 for the mother frailty. In fact, no interval cover zero value of the random effect variance, suggesting that the household and mother frailty are important. For community and village frailty, the 95% confidence intervals cover the zero value of the random effect variance, indicating that the community and village frailty are not important. This confirms the results of Table 3.1, in which household and mother frailty are important, whereas community and village frailty are not.

High household frailty effect indicates that housing condition, socio-economic status and other household level factors are more important than other factors that operate at mother, community or village level. The mother's frailty effect was lower than the household, probably because some of the important maternal variables for childhood mortality have been accounted for in the model, such as maternal education and maternal age at delivery, whereas none of household's variables have been included. It is suggested that household factor variables should be included for further studies.

Similar to the infant mortality model, the estimated parameters in the child mortality models with frailty do not differ from the standard model (Table 3.2).

The general conclusion regarding the sibling and gender factors is that there was no evidence of gender difference reflected as difference in care between boys and girls in Purworejo district, Indonesia that may lead to mortality. This finding is in accord with the previous research (Wahab et al., 2001) and the general trend of the narrowing gaps in many aspects between boys and girls in the Indonesian society (Kevane and Levine, 2003). There is, however, an indication that having brother(s) may lead to higher risk of child mortality.

3.3 Morbidity: surveillance data

Because of its importance, childhood morbidity has been investigated by many researchers from diverse disciplines such as public health, biomedicine and social science. Two common diseases in childhood, diarrhea and respiratory infection, remain to be the most important causes of deaths among children (Rice, Sacco, Hyder and Black, 2000; Black, Morris and Bryce, 2003; UNICEF, 2003). In Indonesia, especially in the CHN-RL area, several studies related to childhood morbidity have been conducted. Machfudz (1998) conducted a study on the effect of morbidity (diarrhea and respiratory infection) on the change of the mid-upper-arm circumference in children under five years of age. Danardono (2000) studied the multilevel effects at community level, household level and individual level for the case of diarrhea disease. Wibowo (2000) evaluated the influence of nutritional status on morbidity (diarrhea and respiratory tract infection) among infants.

We presented the application of EHA for analyzing two common and important childhood diseases, diarrhea and respiratory infection in the CHN-RL surveillance area. We demonstrated the use of various time scales to respond to research questions of interest. As in the previous section, the detail of the analysis in this section has been reported elsewhere by Danardono (2003).

3.3.1 Data, study variables and models

We utilized the CHN-RL morbidity surveillance for this analysis. The surveillance used the *two-week recall questionnaire* to collect information on childhood morbidity at the day of visit and 14 days backward and related variables. This type of questionnaire has been widely used for morbidity records, for instance in the Demographic and Health Surveys (DHS) in many countries, including Indonesia (CBS, NFPCB, MOH and MI, 1998).

The variables of interest are *gender* of the child, *maternal education* and *maternal age* (at the time of illness), *sibling* variables (as in the childhood mortality models in the previous section) and *breastfeeding*. Individual frailty effects as well as environmental and institutional frailty effects are also investigated. To ensure that information on the breastfeeding variable is available, cohort data from February 1995 until June 1998 were used with 2804 children available in the data set.

To analyze the data, we need to construct the data set into counting process style of input `(start, stop], event`. The process is straightforward but tedious, and computer demanding when the data set is large and includes time dependent covariates. Table 3.3 represents the data layout for the morbidity study. The observation column is the information obtained by the two-week recall questionnaire. The start, stop, event columns are constructed by the observation column and visit column. For instance, child with ID

Table 3.3: Data layout for morbidity study. In this example there are 2 children with 2 and 4 visits resulting into 9 spells (intervals with `(start, stop]` and `event`). The event of interest is `1` in the *observation* column. Some observations are split because the occurrence of the event or time-dependent covariate (e.g., weaned)

ID	start	stop	event	observation	visit	weaned
96	96-05-15	96-05-29	0	000000000000000	96-05-29	——
96	96-08-20	96-08-31	1	000000000001111	96-09-03	——
81	96-10-23	96-10-26	1	000111110000000	96-11-06	97-07-31
81	96-10-31	96-11-06	0		96-11-06	97-07-31
81	97-01-31	97-02-14	0	000000000000000	97-02-14	97-07-31
81	97-04-29	97-05-07	1	000000001111111	97-05-13	97-07-31
81	97-07-25	97-07-31	0	000000000011100	97-08-08	97-07-31
81	97-07-31	97-08-04	1		97-08-08	97-07-31
81	97-08-07	97-08-08	0		97-08-08	97-07-31

81 at visit `1996-11-06` was split into two intervals, one ended at `1996-11-26` with event, and one at `1996-11-06` censored. The dates are constructed backwards in time from the visit date. When there are changes in the value of the time-dependent covariate, such as weaned at `1997-03-31`, the observation was split according to the covariate times (e.g., ID 81 at visit `1997-08-08`)

We use the the Andersen-Gill (AG) model, an extension of Cox's model with age time scale, calendar time, and time since weaning. The model assumes independent increments, i.e., the numbers of events in non-overlapping time intervals are independent, given the history, with common baseline hazards for all events.

The AG model specifies intensity process similar to hazard function in the Cox model

$$\lambda(t|\mathbf{Z}(t)) = Y(t)\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}(t)), \tag{3.1}$$

where $\lambda_0(t)$ is the baseline intensity, $\boldsymbol{\beta}$ is unknown regression coefficients, $\mathbf{Z}(t)$ is vector of covariate, possibly time-dependent and $Y(t)$ is zero-one at-risk process. Unlike the Cox model for survival data, $Y(t)$ in the AG model is not absorbed to zero when an event occurs but alternates between zero and one depending on the event process. The purpose of counting process style of input `(start`, `stop]`, `event` mentioned above is to specify the $Y(t)$.

In the analysis we used the AG model with three different time scales, i.e., age, time since the start of the surveillance, and time since weaning.

3.3.2 Age time scale

Respiratory infection and diarrhea, as well as many other childhood diseases are usually age dependent. Choosing age as the time scale does not allow age itself to be in the model, but we can check the dependency by looking at the hazard plot. Figure 3.3 shows the plot of the hazards for both diseases. The hazard plots are smoothed by the *Epanechnikov* kernel, with a bandwidth of 10 months chosen by visual inspection, and plotted over the monthly crude hazard rates (the shaded barplot). The visual inspection is of course not an optimal method for choosing a bandwidth, compared with the method suggested by Andersen et al. (1993), but it is useful enough for exploratory purposes. The cumulative hazards of both diseases are almost linearly increasing. The estimated hazards show that the hazard might be associated with age, and around 12 months of age could be the highest peak of both diseases.

Table 3.4 gives the result for diarrhea. Increasing maternal age seems to be associated with increasing the risk. The *breastfeeding* variable has a rather significant contribution to the model where the *never breastfed children* had the highest risk as compared to the other categories. Maternal education and sibling variables did not show any significant contribution in the model.

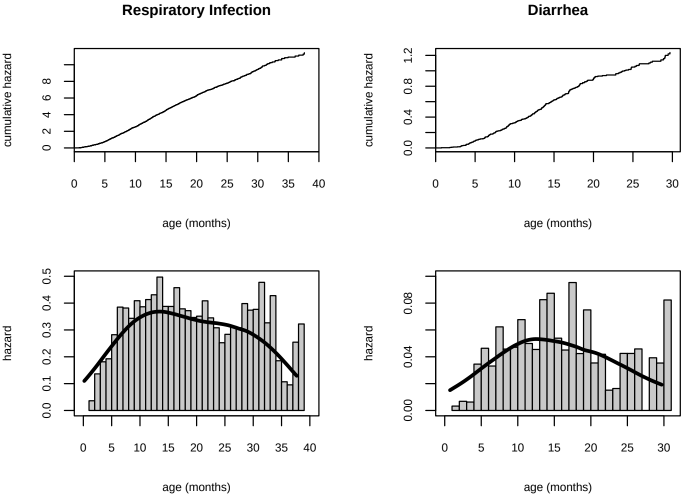


Figure 3.3: The cumulative hazard and hazard plot of childhood respiratory infection and diarrhea by age.

Table 3.4: Hazard model for diarrhea, age time scale

Variables	Relative risk	(c.i.)	p-value LRT[1]	p-value Non-prop[2]
Gender			0.298	
boy	1	(reference)		
girl	1.14	(0.87-1.51)		0.581
Maternal education			0.174	
non-educated	1	(reference)		
educated	1.58	(0.8-3.14)		0.784
Breastfeeding status			0.088	
breastfed	1	(reference)		
weaned	1.09	(0.69-1.73)		0.179
never breastfed	2.07	(1.04-4.11)		0.260
Maternal age (years)			<0.001	
15-19	1	(reference)		
20-24	1.40	(0.75-2.60)		0.555
25-29	1.70	(0.95-3.03)		0.247
30-34	1.14	(0.59-2.22)		0.731
35+	1.66	(0.86-3.21)		0.383
Older sibling			0.640	
none	1	(reference)		
brother	0.84	(0.57-1.24)		0.902
sister	0.91	(0.62-1.33)		0.846
Younger sibling			0.988	
none	1	(reference)		
brother	1.12	(0.29-4.39)		0.630
sister	0.99	(0.27-3.67)		0.991

[1] Likelihood ratio test [2] Non-proportionality test, global p-value=0.89

The frailty effects of this hazard model for diarrhea are all significant, with the value of 1.273, 1.229, 1.237, 0.614, 0.350 for individual, mother, household, community and village frailty, respectively. The estimated coefficients in the frailty models are only slightly different to the estimated coefficients of the standard model (Table 3.4), which may not give any further important information. However, the significant frailty effect of these frailty models indicate the existence of unobserved heterogeneity in the individual, mother, household, community and village groups which may need to be investigated further.

For the respiratory infection, the maternal age has a similar pattern to the diarrhea models as well as for the sex and sibling variables. Contrary to the diarrhea model, maternal education gave significant contribution to the model whereas the breastfeeding variables did not. The frailty effects for the respiratory infection model were also found to be important in the models.

For both hazards models of respiratory infection and diarrhea, there is no evidence of non-proportionality, as indicated by the global p-values test for non-proportionality (large values) and the p-values for each coefficient in both models. All necessary interactions, such as maternal education and breastfeeding, have also been checked and taken care of.

3.3.3 Calendar time

We used other time scales than age to allow age as a time dependent covariate in the model. One possible choice is the time since the start of the surveillance (February 1995).

Figure 3.4 shows the cumulative hazards and hazard as a function of time since the start of the surveillance where time is converted back to a calendar time. Against time, the hazard of respiratory infection is always higher than diarrhea. The highest peak of respira-

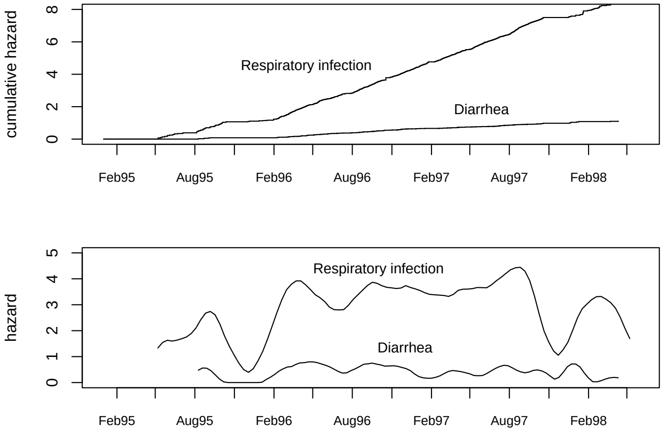


Figure 3.4: The cumulative hazards and hazards plot of childhood respiratory infection and diarrhea by calendar time.

tory infection and diarrhea incidence seemed to be in April-June, the transition period from the rainy to the dry season; and in September-October, the transition from the dry to rainy season. There was also a long dry season in 1997 and an economic crisis that might have caused the peak incidence in that year.

Table 3.5 shows the hazards model for respiratory infection. The *children's age* variable is significantly associated with the risk of developing respiratory infection. The highest risk for respiratory infection is in the `6-23` (months) age group. The conclusion is the same for *maternal education* and *maternal age* as in the model using age time scale. The other variables have a similar pattern to the models using age as the time scale. The pattern is also similar for the diarrhea models.

Also similar to the age time scale models, introducing frailty did

Table 3.5: Hazards model for respiratory infection, calendar time

Variables	Relative risk (c.i.)		p-value LRT[1]	p-value Non-prop[2]
Gender			0.984	
boy	1	(reference)		
girl	0.99	(0.91-1.10)		0.135
Age of the child (months)			<0.001	
0-5	1	(reference)		
6-23	1.83	(1.61-2.07)		0.267
24+	1.51	(1.21-1.90)		0.722
Maternal education			0.013	
no education	1	(reference)		
6 yrs of education	1.33	(1.04-1.70)		0.168
9 yrs of education	1.29	(0.99-1.69)		0.145
12 yrs of education	1.42	(1.09-1.85)		0.259
Maternal age (years)			<0.001	
15-19	1	(reference)		
20-24	1.29	(1.05-1.58)		0.303
25-29	1.30	(1.04-1.61)		0.367
30-34	1.15	(0.92-1.45)		0.461
35+	1.32	(1.04-1.67)		0.656
Breastfeeding status			0.674	
breastfed	1	(reference)		
weaned	1.01	(0.87-1.18)		0.725
never breastfed	0.84	(0.55-1.28)		0.595
Older sibling			0.641	
none	1	(reference)		
brother	0.98	(0.85-1.12)		0.560
sister	0.94	(0.82-1.08)		0.458
Younger sibling			0.480	
none	1	(reference)		
brother	0.98	(0.57-1.68)		0.661
sister	1.32	(0.85-2.05)		0.922

[1] Likelihood ratio test [2] Non-proportionality test, global p-value=0.93

not change the estimated coefficients for both respiratory infection and diarrhea models, but the frailty variance was quite significant indicating unobserved heterogeneity in the data. Neither model violates the proportionality assumption of the Cox proportional hazard model according to the non-proportionality test.

3.3.4 Time since weaning

The protective effect of breastfeeding for childhood illness is well known and has been investigated by many authors, see for example Bhandari, Bahl, Mazumdar, Martines, Black, Bhan and Infant Feeding Study Group (2003) and references therein. In this section, the aim is to demonstrate the use of *time since stop breastfeeding* as an alternative time scale, for investigating the effect of breastfeeding on childhood morbidity. The breastfeeding definition in this section is simply based on the questionnaire on health status, breastfeeding and feeding practice and does not include breastfeeding pattern, such as exclusive breastfeeding and frequency of breastfeeding. The percentage of breastfed children is quite high in the surveillance area, about 98%, which is similar to the national figure of 96% (CBS et al., 1998). The median duration of breastfeeding is relatively long at 24.1 months which is also close to the national figure of 23.9 months.

The weaned age or the duration of breast feeding is one of the variables of interest. This variable is a time independent covariate that is fixed since the weaned time. Age of the child is also included as a time dependent covariate.

Table 3.6 gives the hazards model for respiratory infection. The *weaned age* is significant in the model. Although the differences of the effects between weaned age category are not huge, the longer weaned age seems to give a protective effect against respiratory infection.

Contrary to the previous models (with age and calendar time scale), the effect of *maternal education* is weak and leads to a different direction. The *maternal age* has similar pattern to the previous models. As with the previous models, there is no evidence of *gender* and *sibling* effect in this model.

The frailty effects are significant but do not change the general conclusion of the model (the estimated coefficients). In this model, there is no indication of violating the non-proportionality assumption.

The hazards model for diarrhea generally gives similar results as for respiratory infection. Here, the results are presented only for the weaned age. The variable has fewer categorizations than for respiratory infection because of the fitting problem. The relative risks with confidence intervals are 0.97 (0.25-3.76), 0.53 (0.13-2.13) for weaned age group `6-12` and `12+` months (the reference is `0-5` months) and has a p-value (LRT) of 0.607.

The frailty effects do not change the coefficient estimation of the hazards model for diarrhea. In fact, no frailties effects for diarrhea are significant. It may really show that there are no unobserved factors for the risk of diarrhea or no difference in risk in these groups or clusters (individual, mother, household, community and village). However, it is also possible that the number of observations is not large enough to show the frailty effects.

In general, the analysis concludes that children aged 6-23 months, or aged around one year of age, are prone to develop respiratory infection and diarrhea and there is a pattern of seasonality in both diseases. Maternal education is important. Surprisingly, the risk of the children developing the diseases are higher for the higher educated mothers. As in the mortality study, there is no evidence of gender and sibling's effect.

Table 3.6: Hazards model for respiratory infection, time since weaning

Variables	Relative risk (c.i.)		*p*-value LRT[1]	*p*-value Non-prop[2]
Gender			0.123	
boy	1	(reference)		
girl	1.22	(0.95-1.58)		0.925
Weaned age(months)			0.040	
0-4	1	(reference)		
5-6	0.48	(0.18-1.32)		0.904
7-12	0.97	(0.55-1.74)		0.314
13+	0.58	(0.33-1.04)		0.431
Age of the child (months)			0.684	
0-5	1	(reference)		
6-23	1.41	(0.59-3.34)		0.747
24+	1.35	(0.52-3.51)		0.516
Maternal education			0.341	
no education	1	(reference)		
6 yrs of education	0.84	(0.41-1.7)		0.113
9 yrs of education	0.98	(0.47-2.12)		0.191
12 yrs of education	0.88	(0.42-1.86)		0.197
Maternal age			0.018	
15-19	1	(reference)		
20-24	2.44	(1.03-5.76)		0.851
25-29	3.23	(1.36-7.68)		0.551
30-34	2.32	(0.94-5.74)		0.564
35+	2.36	(0.93-5.97)		0.800
Older sibling			0.567	
none	1	(reference)		
brother	0.84	(0.59-1.2)		0.794
sister	0.84	(0.58-1.21)		0.688
Younger sibling			0.495	
none	1	(reference)		
brother	0.77	(0.36-1.65)		0.507
sister	1.26	(0.73-2.16)		0.633

[1] Likelihood ratio test [2] Non-proportionality test, global *p*-value=0.86

3.4 Morbidity: trial data

Deficiencies of iron and zinc often coexist and cause growth faltering, delayed development and increased morbidity from infectious diseases during infancy and childhood (Lind, 2004, Paper V). Therefore, combined iron and zinc supplementation may be a logical prevention strategy.

To investigate the effect of the supplementations, a community-based, randomized, double-blind, controlled trial, the ZINAK study, was conducted from July 1997 to May 1999 in the CHN-RL area, Purworejo, Indonesia. The subjects are different to the children in the surveillance morbidity discussed in the previous section.

This section demonstrates the use of EHA for morbidity analysis in the ZINAK data. Unlike the morbidity analysis in the previous section, here, we have continuous data collection in which various analyses rather than only AG-model are possible to be performed. We considered respiratory infection as the event of interest. Together with infant growth analysis in the next section, this section serves as a background problem for Chapter 5.

3.4.1 Data, study variables and models

The ZINAK study was conducted from July 1997 to May 1999 in the CHN-RL surveillance area, Purworejo, Indonesia. Healthy and singleton infants, aged less than six months were recruited. After assessing their eligibility, 680 infants were randomized into one of four treatments: iron, zinc, iron+zinc or placebo from 6 to 12 months of age (180 days of supplementation). More detailed description of the design and data collection is reported by Lind (2004). There are several outcomes of interest, biochemical outcomes (iron and zinc concentration in the blood), infants growth (anthropometry), infants development (mental, psychomotor development) and morbidities. Here, we consider respiratory infection as the outcome of interest.

Morbidity information was obtained by visitation every third day. Field workers asked the parents or guardians regarding the compliance to supplementation as well as information on symptoms of illness for the day of visit and for the two days preceding the visit.

Among 680 infants, 666 completed supplementations and some of them dropped out. It may be necessary to consider the drop-out in the analysis since all of them were related to the supplementation as reported by Lind (2004). However, at this moment we analyze the completed records only according to *intent-to-treat* analysis. Covariates under consideration, other than the treatment itself, are gender and maternal education.

We used the AG model with age as the time scale as in the previous section (Equation (3.1)). Additionally, we used gap-time or sojourn time also as an alternative time scale. The gap-time is defined as the time since entry or previous event. When both models give similar results, we can safely assume a renewal process and consider a constant baseline hazard.

As in the previous section, we may actually use calendar time as well since morbidity may have a strong seasonal pattern. However given the rather short period of observation time (six months) and that most of the children entered the study at almost the same time, using calendar time and age is almost identical. However, when we want to model the morbidity with growth, which depends on age rather than calendar time, the age time scale has a clear advantage to calendar time.

3.4.2 Results

Tables 3.7 and 3.8 give the result of hazard model using the AG model and gap-time model. They are actually quite similar in their

Table 3.7: Hazards model for respiratory infection using the Andersen Gill model, ZINAK study

Variables	Risk ratio	(c.i.)	p-value LRT[1]	p-value Non-prop[2]
Gender			0.044	
boy	1	(reference)		
girl	0.91	(0.83-1.00)		0.308
Supplementation			0.411	
placebo	1	(reference)		
zinc	1.00	(0.88-1.14)		0.805
zinc+iron	0.91	(0.79-1.03)		0.235
iron	0.97	(0.85-1.11)		0.723
Maternal Education			<0.001	
no-education	1	(reference)		
6 years	0.84	(0.64-1.10)		0.133
9 years	0.70	(0.53-0.92)		0.427
12 years or more	0.46	(0.29-0.75)		0.102

[1] Likelihood ratio test [2] Non-proportionality test, global p-value=0.177

risk ratio and p-value of the likelihood ratio test. Assuming constant baseline hazards will give the same result. The raw and smoothed hazard function in Figure 3.5 also indicated a constant hazard during period of 6 to 12 months of age.

Looking at the estimates, there is no pronounced effect of the supplementation to respiratory infection which confirms the result by Lind (2004) in which Poisson regression was used. This result also reiterates the importance of maternal education as it has been found in the respiratory infection models using surveillance data (Section 3.3). Here, the direction of the maternal education is different to that of surveillance data. Higher education seemed to have protective effect on respiratory infection. The infants' gender was rather significant with girls having a lower hazard than the boys.

Table 3.8: Hazards model for respiratory infection using the gap-time model, ZINAK study

Variables	Risk ratio	(c.i.)	p-value LRT[1]	p-value Non-prop[2]
Gender			0.051	
boy	1	(reference)		
girl	0.91	(0.83-1.00)		0.014
Supplementation			0.474	
placebo	1	(reference)		
zinc	1.01	(0.89-1.15)		0.172
zinc+iron	0.91	(0.80-1.04)		0.791
iron	0.98	(0.86-1.11)		0.652
Maternal Education			<0.001	
no-education	1	(reference)		
6 years	0.85	(0.65-1.12)		0.484
9 years	0.72	(0.54-0.95)		0.176
12 years or more	0.50	(0.31-0.80)		0.883

[1] Likelihood ratio test [2] Non-proportionality test, global p-value=0.101

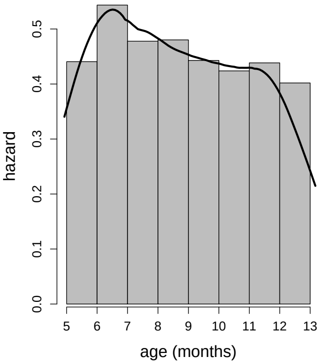


Figure 3.5: Raw and smoothed hazard plot of childhood respiratory infection by age.

The other purpose of this analysis, aside from demonstrating the application of EHA, is to give a background for the problem of analyzing EHA together with longitudinal measurements in Chapter 5. It is well known that nutrition, growth and morbidity are closely related (Scrimshaw, 2003). Therefore, evaluating supplementation on both growth and morbidity simultaneously may give less bias than analyzing the two outcomes separately. Although it also has been reported briefly that anthropometrical status was not associated with the incidence of infectious disease (Lind, 2004), a more careful analysis may be needed.

3.5 Infant growth

Infant growth indicators such as weight, length, knee-heel, mid-upper arm circumference are another outcome of interest collected in the ZINAK study. Obviously, the type of outcomes is not a time-to-event data but ordinary continuous data. We presented the use of LDA to analyze such data, taking weight as the outcome of interest. Also, together with the morbidity analysis in the previous section this section serves as a background problem for Chapter 5.

Measurements of the weight were performed every month. Weight measurements before the period of trial were also available for most of the children. Figure 3.6 shows the children's weight by age with smoothing lines. During the trial period from 6 to 12 months of age, a linear model for this weight growth curve may be sufficient. However, weight growth is very individually developed in which the between individual variance is usually large. Therefore, employing the linear random effects model reviewed in Section 2.3.2 is more suitable to the weight data than the ordinary linear model.

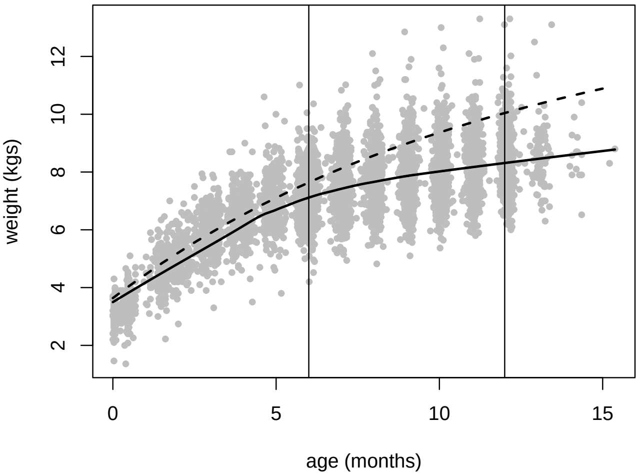


Figure 3.6: The children's weight across age. The greyed points denote the actual measurements of weight; the line denotes the smoothing splines of the weight measurements; the dashed line denotes the reference population (CDC 2000 growth charts); and the two vertical lines denote the starting and ending point of the trial.

Table 3.9: Growth curve model for weight using random effect and ordinary linear model, ZINAK study

Variables	Random effect model		linear model	
Intercept	6.37	(5.88,6.86)	6.38	(6.11, 6.65)
Age	0.17	(0.17,0.18)	0.17	(0.15, 0.19)
Gender				
boy		(reference)		
girl	-0.54	(-0.68 ,-0.40)	-0.54	(-0.61,-0.48)
Supplementation				
placebo		(reference)		
zinc	0.02	(-0.18 , 0.22)	0.08	(-0.01, 0.16)
zinc+iron	0.01	(-0.19 , 0.21)	0.02	(-0.07, 0.10)
iron	0.01	(-0.19 , 0.21)	0.03	(-0.06, 0.12)
Maternal Education				
no-education		(reference)		
6 years	0.20	(-0.28 , 0.68)	0.19	(-0.02, 0.40)
9 years	0.31	(-0.18 , 0.79)	0.30	(0.09, 0.51)
12 years or more	0.26	(-0.41 , 0.94)	0.27	(-0.03, 0.57)
Illness days	-0.53	(-0.64 ,-0.41)	-1.12	(-1.57,-0.67)
Random effect				
sd(Intercept)	0.993	(0.923,1.064)		
sd(Age)	0.065	(0.061,0.070)		
corr(Intercept,Age)	-0.617	(-0.860,-0.430)		

The model for weight is

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \boldsymbol{\epsilon}_i, \qquad & i = 1, \ldots, N, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}), \qquad & \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \end{aligned} \tag{3.2}$$

where $\mathbf{y}_i$ is the weight measurements on child i and N is the number of children, $\mathbf{b}_i$ is vector of random effects, $\mathbf{X}_i$ and $\mathbf{Z}\mathbf{b}_i$ are the covariates for the fixed and random effects, respectively.

Table 3.9 shows the results of fitting the weight models using a random effects model and also the ordinary linear model for comparison. The *age* and *illness-days* covariates are measured as continuous variables while the rest are categorical. Illness days is the number of days with illnesses (symptoms) from the previous measurements time up to the current measurement time, as a proxy variable for the effect of duration of illness.

The random effects model has two parts, the fixed part (upper part of the column variables) and the random part (the lower one). First, we look at the random effects which correspond to the standard deviation and correlation of *intercept* and *age* (the $\mathbf{b}_i$ in model (3.2)). They were found to be significantly different from zero, as indicated by their intervals which do not include zero. The result confirms the assumption that weight growth is quite individually developed.

Now we look at the fixed part ($\boldsymbol{\beta}$ in model (3.2)) and compare the estimates with that of the ordinary linear model. The estimated coefficients of the two models are quite similar except for the *illness-days*. The confidence intervals from the random effect model are generally wider than that from the ordinary linear model. There seemed to be no effect of supplementation on the weight. The pronounced effects were *gender*, *age* and *illness-days*. We have check some interactions as well and we found that there was no interaction between *supplementation* and *illness-days*.

As comparisons, we also performed two alternative analyses for the weight longitudinal measurements. The first one is an analysis with WAZ (weight-for-age z-score) instead of weight. The WAZ is a standardized value of the weight compared to a reference population. We used the CDC 2000 reference population (Kuczmarski, Ogden and Guo, 2002) which was also used by Lind (2004) (see the dashed line in Figure 3.6). The *age* and *gender* variables were important, similar to the weight random effect model of Table 3.9, but the direction of the estimated regression coefficients was reversed. The estimated 95% confidence intervals were (-0.22, -0.206) and (0.00, 0.33) for *age* and *girl*, respectively. This indicates growth decreasing compared to the growth of the CDC 2000 and the boys seemed to suffer more than the girls. There is no different in conclusion for the supplementation, maternal education and illness days.

The second one is an analysis using weight velocity. The weight velocity for a certain age of individual is the weight difference between the current weight and the previous measurement weight divided by the length of time from the previous measurement age to current age. We used the ordinary linear model as the random effect part did not show any significant contributions. The age and illness days still show a large effect as in the weight models. The gender effect, however, disappeared. As in the weight models, supplementation did not show any significant effect in this weight velocity model.

In conclusion, there is a general growth decrease for children in the study compared to the standard reference population, but the supplementation did not seem to affect the growth. It is also of interest to investigate the growth model in relation to time-to-event morbidity data. We will discuss this problem in Chapter 5.

3.6 Remarks

We have demonstrated the application of EHA and LDA to analyze data from childhood health studies. There are two points of methodological interest emerging from the applications.

In EHA, sometimes we face more than one competing time scale. For instance, we may use calendar time instead of age in the morbidity model of Section 3.3. Age-period or age-period-cohort model is another situation in which more than one time scale is involved. The problem of multiple time scales will be discussed in the next chapter.

Important statistical issues in the ZINAK study is that the outcomes of interest may actually interact with each other and analyzing them separately may give biased results. Specifically, the interest is on the joint analysis of time-to-event and longitudinal measurements outcomes. Comparison of approaches and further analysis of ZINAK respiratory infection and growth data will be presented in Chapter 5.

Chapter 4

Multiple Time Scales

4.1 Introduction

Time is indispensable in event history analysis. Although time may be just a proxy measure for other influences of the events (Berzuini and Clayton, 1994b), time is the most readily available measurement and easy to utilize for comparison and generalization. For example, in epidemiology, age is the most often used time scale since it reflects cumulative damage that causes mortality, whereas, in clinical studies, time since diagnosis may be more important. This chapter considers the problem of choosing an appropriate baseline time scale and modeling dual time scales.

The choice of time scale is driven by the research question of the study. However, in the absence of knowledge about the importance of time scales, we may have to consider all relevant time scales. In an epidemiological surveillance study, it is common to perform an exploratory study to identify new emerging risk factors. One way of exploring the factors is by investigating several relevant time scales. In general, the choice of relevant time scales in epidemiology or observational studies is more difficult than in clinical studies (Liestøl and Andersen, 2002).

Farewell and Cox (1979) and Oakes (1995) suggested to choose a basic time scale that accounts for as much as the variation as possible. Duchesne (1999) and Duchesne and Lawless (2000) introduced the concept of *ideal time scale*. However, their focus is on the *usage variable* (such as mileage, asbestos exposure, etc.), as the other scale rather than the multiple origins problem, as considered in this thesis. Multiple time scales have been considered in the multistate model as well. Jones and Crowley (1992) and Commenges (1999) considered the problem of multiple time scales under the Markov and semi-Markov models. Ng and Cook (1997) developed a random effects model that includes piecewise constant formulations. Andersen and Keiding (2002) suggested a practical approach to choosing a basic time scale in the Cox model.

The piecewise constant hazards and discrete time models are the usual approaches to the multiple time scales problem, if we want to treat multiple time scales symmetrically (Keiding, 1990; Berzuini and Clayton, 1994b). Those approaches utilize the relation between Poisson regression and Cox's proportional hazards model. Efron (2002) considered the discrete time approach to develop a *two-way* proportional hazards model and decomposed the hazards multiplicatively for a dual time scales problem.

In the Cox model, other time scales (than the basic time scale) can be considered as a defined time-dependent covariate (see Section 2.4.1). Therefore, Cox models with a time-dependent approach, such as a time-dependent covariate and time-dependent strata, can be used for multiple time scales modeling.

In this chapter, procedures to choose a basic time scale in Cox's regression model are proposed. For the dual time scales problem, the connection between piecewise constant hazards and the time-dependent approach is discussed. Quantitative comparisons are performed through simulation.

4.2 The choice of relevant time scales

The multiple time scales problem considered here is basically a multiple time origins problem with time equal to ordinary clock time. The nature of the problem is different from the usual multivariate survival such as bivariate survival in twin studies or pairs of human organ studies. In the multiple time origins problem, see the Lexis diagram in Figure 4.1(a), movement of time scale pairs is in the same direction (a line with slope 1) (Keiding, 1990). When a subject dies, for instance, both movements for that subject stop. In twin studies, a pair of twins may have different paths, if one dies the other may still continue the path.

Figure 4.1 shows the life line in a Lexis diagram for one subject and its corresponding separate time scales. Usually the time on the abscissa (T_1) represents calendar time, life length measured from the "zero" calendar date (the birth of Christ); whereas time on the ordinate (T_2) represents age, life length measured from the subject's birthdate. Another example is in a clinical study, where T_1 represents age and T_2 represents time-since-diagnosis. As we can see from the figure, both time scales stop at the same event time (the dashed lines) at a certain reference time, but their origins are different. The problem is choosing the most relevant time scale as baseline.

There is no regression coefficient estimated for the basic time scale. Therefore, a time variable whose effect is of interest should not be used as the basic time scale (Andersen and Keiding, 2002). However, the time variable with suspiciously irregular effect, which is difficult to model parametrically via a time-dependent covariate, may be chosen as the basic time scale.

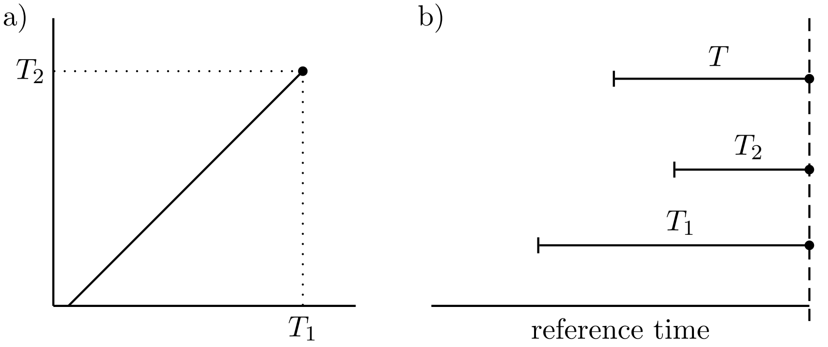


Figure 4.1: (a) Lexis diagram and (b) separate scale

The guideline may be useful enough in practice, yet there is another situation when a more formal procedure in choosing a time scale is needed. When there is a suspicion about the erroneously specified time origin we may need a formal procedure to examine the observed time scales. We call a procedure to deal with the problem an *erroneous scale* procedure, henceforth.

The erroneous scale model assumes a data generating mechanism as in Figure 4.1. The hazard function of a true but unobserved duration T is modeled as a Cox model

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta} \mathbf{Z}_i(t)), \quad t > 0, \tag{4.1}$$

where $\lambda_i(t)$ is the baseline hazard function for subject i.

Several alternative time origins might be observed, resulting in several time scales (durations), e.g., T_1 and T_2 in Figure 4.1. In a real situation, the true duration T may be the time since onset until the event of interest which is not observable, and the alternative durations T_1 and T_2 are age and time since diagnosis, respectively. We are interested in choosing one most relevant time scale as a surrogate of the true time scale.

The Cox model with alternative time scales can be specified as

$$\lambda_i(t) = \lambda_0(t + \delta_i) \exp(\boldsymbol{\beta} \mathbf{Z}_i(t + \delta_i)), \quad t > 0, \tag{4.2}$$

where δ_i represents the difference or delay between the true origin and the alternative origins for subject i. For example, δ_i is the duration from onset until diagnosis.

In this situation we may not have a proportional hazards model any longer since δ_i varies between individuals. Therefore, when we observe only the alternative time scales, a simple procedure to investigate whether the time scale is appropriate or not is by examining the proportional hazards assumption.

We can write the hazard $\lambda_0(t + \delta_i)$ as $\lambda_0(t_0)W_i$, separating the baseline hazards and the subject-specific factor, if we assume the Gompertz hazard function (Liestøl and Andersen, 2002). Model (4.2) then is a Cox model with frailty (Section 2.2.5),

$$\lambda_i(t) = \tilde{\lambda}_0(t) W_i \exp(\boldsymbol{\beta} \mathbf{Z}_i(t + \delta_i)), \quad t > 0, \tag{4.3}$$

where W_i is the random effects or frailty variable as a function of δ_i. In this situation we may estimate a frailty effect, for instance by assuming that W_i is gamma distributed with mean 1 and variance ω. Therefore, another procedure to examine the time scales is by examining the frailty effects.

However, when those procedures do not seem to reveal the most relevant time scale, and there is scientific reason that the time scales are all important, we may model multiple time scales simultaneously. We discuss this problem for the case of dual time scales in the next section.

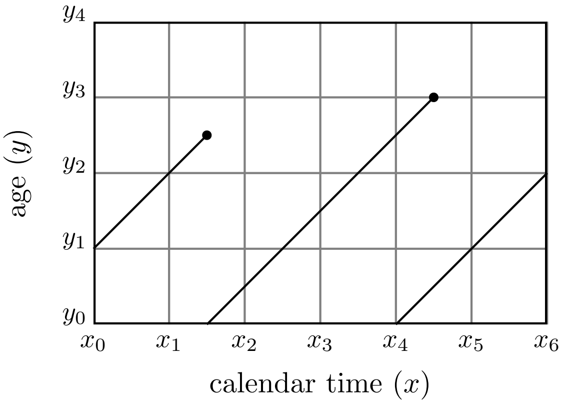


Figure 4.2: Hypothetical event history data on a Lexis diagram. The lines represent the observed follow-up time by age and calendar time (period); the dots represents the event of interest (deaths, diseased)

4.3 Modeling dual time scales

We will discuss the multiple time scales problem for the case of dual time scales such as age and calendar time (period). Figure 4.2 represents typical dual time scales event history data on a *Lexis* diagram.

The general aim is to model the hazards as a function of age y, calendar time x and covariate $\mathbf{Z}$ which may also depend on y and x. Let $\mu(x, y)$ be the hazard function at period x and age y. Generalizing from the single time scale, the Cox proportional hazard model for dual time scales is

$$\mu(x, y \mid \mathbf{Z}) = \mu_0(x, y) \exp(\boldsymbol{\beta}\mathbf{Z}), \tag{4.4}$$

where $\mu_0(x, y)$ is the baseline hazard function at period x and age y common to all individuals. Three approaches are considered here to model (4.4), i.e., the piecewise constant hazards, time-dependent strata and time-dependent covariate methods.

4.3.1 Piecewise constant hazards

In the piecewise constant hazards model we assume that the hazard function $\mu(x, y)$ is piecewise constant across the Lexis plane. Technically, the Lexis plane is divided into sufficiently small rectangles such that constant hazard function μ can be reasonably assumed in each rectangle. Let u_i be the total exposure time in a rectangle and d_i be the number of events (0 or 1) for individual i, then the contribution of individual i to the likelihood is

$$L_i(\mu) = (\mu)^{d_i} \exp(-\mu u_i), \quad i = 1, \ldots, n, \tag{4.5}$$

in this specific rectangle. To assess other effects on the hazard we may specify $\mu \exp(\mathbf{Z}\boldsymbol{\beta})$ instead of only μ, where $\mathbf{Z}$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. Although any functional form of $\mathbf{Z}$ and $\boldsymbol{\beta}$ is possible, the log-linear form $\exp(\boldsymbol{\beta}\mathbf{Z})$ is convenient.

Let the Lexis plane, as in Figure (4.2), be divided into smaller rectangles

$$\Omega_{(r,s)} = \{(x, y) : x \in [x_{r-1}, x_r) \text{ and } y \in [y_{s-1}, y_s)\},$$

$r = 1, 2, \ldots, R$, $s = 1, 2, \ldots, S$; $d_{i(r,s)}$ and $u_{i(r,s)}$ be the number of observed events and time spent (exposure time) in each $\Omega_{(r,s)}$ for individual i, $i = 1, \ldots, n$. The likelihood for the piecewise constant hazards model (4.5) for all individuals and over the lexis grid Ω is

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}) = \prod_{r=1}^{R} \prod_{s=1}^{S} \prod_{i=1}^{n} \left[\left(\mu_{rs} e^{\boldsymbol{\beta}\mathbf{Z}_i}\right)^{d_{i(r,s)}} \exp(-\mu_{rs} e^{\boldsymbol{\beta}\mathbf{Z}_i} u_{i(r,s)}) \right], \tag{4.6}$$

where μ_{rs} is the baseline hazard in $\Omega(r, s)$.

It is possible to assess the effects of time (age and calendar time) on the hazard by assuming a multiplicative decomposition $\mu_{rs} = \lambda_s \gamma_r$.

4.3.2 Time-dependent approaches

In the single time scale situation, the partial likelihood used in the Cox proportional hazards model to estimate the regression coefficients can be interpreted as a profile likelihood obtained from a piecewise constant hazards likelihood maximized to certain nuisance parameters and allowing the width of the time intervals approaching zero (Johansen, 1983; Clayton, 1988). This procedure does not work in the dual time scale situation due to the lack of smoothness of the maximum likelihood baseline rate estimates (Keiding, 1990; Berzuini and Clayton, 1994b). Efron (2002) was able to construct a genuine two-way proportional hazards model by considering discrete time scales.

An alternative approach is to let the partition of one time scale interval be fixed as the partition in the other direction gets finer and finer. In the limit, we get two different solutions, depending on which partition is kept fixed.

We consider the likelihood for the piecewise constant hazards model of Equation (4.6). Now, given $\boldsymbol{\beta}$, the μ_{rs} may be separately estimated as follows. Looking at specific values of r and s and suppressing the dependence of them, and taking logs gives

$$\ell_{rs} = \sum_{i=1}^{n} \left[d_i \log \mu + d_i \boldsymbol{\beta} \mathbf{Z} - \mu \exp(\boldsymbol{\beta} \mathbf{Z}_i) u_i \right]. \tag{4.7}$$

By equating the derivative of (4.7) wrt μ to zero, we get $\hat{\mu}_{rs}(\boldsymbol{\beta})$:

$$\hat{\mu}_{rs}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} d_{i(r,s)}}{\sum_{i=1}^{n} u_{i(r,s)} \exp(\boldsymbol{\beta} \mathbf{Z}_i)}, \quad r = 1, \ldots, R; \; s = 1, \ldots, S. \tag{4.8}$$

By replacing μ_{rs} in (4.6) by (4.8), taking logarithms and simplifying, we get the profile log likelihood

$$\ell_p(\boldsymbol{\beta}) \propto \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{n} d_{i(r,s)} \log\left(\frac{e^{\boldsymbol{\beta}\mathbf{Z}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{\beta}\mathbf{Z}_j} u_{j(r,s)}}\right). \qquad (4.9)$$

The time-dependent strata approach

We proceed with the approach with a fixed period (calendar time) x scale, i.e., we keep R in (4.9) fixed.

Now let $\omega = y_s - y_{s-1}$ be the constant width of the time intervals on the y scale. When $S \to \infty$ ($\omega \to 0$), $d_i(r,s)$ and $u_i(r,s)$ will become

$$d_i(r,s) = \begin{cases} 1 & \text{if an event occurs for individual } i \text{ in } \Omega(r,s), \\ 0 & \text{otherwise,} \end{cases}$$

$$u_i(r,s) \approx \begin{cases} \omega & \text{if individual } i \text{ is observed in } \Omega(r,m), \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$Y_i(r,s) = \begin{cases} 1 & \text{if individual } i \text{ is observed in } \Omega(r,m), \\ 0 & \text{otherwise.} \end{cases}$$

The profile likelihood (4.9) then becomes

$$\begin{aligned} \ell_p(\boldsymbol{\beta}) &\approx \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{n} d_{i(r,s)} \log\left(\frac{e^{\boldsymbol{\beta}\mathbf{Z}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{\beta}\mathbf{Z}_j} \omega Y_{j(r,s)}}\right) \\ &= \sum_{r}\sum_{s}\sum_{i} d_{i(r,s)} \log\left(\frac{e^{\boldsymbol{\beta}\mathbf{Z}_i}}{\sum_{j} e^{\boldsymbol{\beta}\mathbf{Z}_j} Y_j(r,s)}\right) \\ &\quad -\sum_{r}\sum_{s}\sum_{i} d_{i(r,s)} \log(\omega) \end{aligned}$$

$$\propto \sum_r \sum_s \sum_i d_{i(r,s)} \log \left(\frac{e^{\boldsymbol{\beta} \mathbf{Z}_i}}{\sum_j e^{\boldsymbol{\beta} \mathbf{Z}_j} Y_j(r,s)} \right) \tag{4.10}$$

removing the terms independent of $\boldsymbol{\beta}$. Since the $d_i(r,s)$ has values 1 only at the event times, the contributions to the likelihood are only at the event times. Therefore the denominator of the log part is actually a sum over the *risk set* given r. The profile likelihood can be written as

$$\ell_p(\boldsymbol{\beta}) = \sum_{r=1}^{R} \sum_{i \in D_r} \log \left(\frac{\exp\left(\boldsymbol{\beta} \mathbf{Z}_i\right)}{\sum_{j \in R_r(y_i)} e^{\boldsymbol{\beta} \mathbf{Z}_j}} \right), \tag{4.11}$$

where D_r is the event set and $R_r(y_i)$ is the risk sets at y_i, given r. In Figure 4.2, event set is all lines with dots, and the risk set is all lines that intersect the horizontal line crossing each dot (the event times y). The profile likelihood (4.11) corresponds to the partial likelihood of Cox's proportional hazards model with basic time scale age y and time dependent strata on the time scale x.

Time-dependent covariate approach

Assuming a multiplicative model for the baseline hazard function,

$$\mu_{rs} = \lambda_s \gamma_r, \quad r = 1, \ldots, R; \; s = 1, \ldots, S, \tag{4.12}$$

we get a slightly different profiling procedure, leading to the time-dependent covariate approach. The log likelihood becomes

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{s=1}^{S} \sum_{r=1}^{R} \sum_{i=1}^{n} \left[d_{i(r,s)} \log\left(\lambda_s \gamma_r e^{\boldsymbol{\beta} \mathbf{Z}_i}\right) - \lambda_s \gamma_r e^{\boldsymbol{\beta} \mathbf{Z}_i} u_{i(r,s)} \right]. \tag{4.13}$$

Given $\boldsymbol{\beta}$ and $\gamma_r, r = 1, \ldots, R$, maximizing (4.13) with respect to $\lambda_1, \ldots, \lambda_S$ is straightforward. The solution is

$$\hat{\lambda}_s = \frac{\sum_r \sum_i d_{i(r,s)}}{\sum_r \gamma_r \sum_i e^{\boldsymbol{\beta}\mathbf{Z}_i}}, \quad s = 1, \ldots, S. \tag{4.14}$$

Substituting (4.14) into (4.13) and simplifying by removing the terms independent of $\boldsymbol{\beta}$ and γ gives the profile likelihood

$$\ell_p(\boldsymbol{\gamma}, \boldsymbol{\beta}) \propto \sum_{i=1}^{n} \sum_{r=1}^{R} \sum_{s=1}^{S} d_{i(r,s)} \log \left(\frac{\gamma_r \exp \boldsymbol{\beta}\mathbf{Z}_i}{\sum_{t=1}^{R} \gamma_t \sum_{j=1}^{n} e^{\boldsymbol{\beta}\mathbf{Z}_j} u_{j(t,s)}} \right). \tag{4.15}$$

We proceed with this derivation in a similar manner to that of the case with time dependent strata. When $S \to \infty$ or $\omega \to 0$, $d_{i(r,s)}$ and $u_{i(r,s)}$ becomes the event indicator and at risk indicator at time s. The summation over all individuals i becomes the summation over the event times $i \in D$. At the denominator of the log part, summation will be determined only at the event times s, since all other terms will vanish by the definition of the event indicator $d_{i(r,s)}$. Similarly, in the denominator the summation will be over the risk set $R(y_i)$. The summation over γ_t will also be completely determined by $j \in R(y_i)$. The profile likelihood becomes

$$\ell_p(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i \in D} \sum_{r=1}^{R} \log \left(\frac{\gamma_r \exp\left(\boldsymbol{\beta}\mathbf{Z}_i\right)}{\sum_{j \in R(y_i)} \sum_m \gamma_m e^{\boldsymbol{\beta}\mathbf{Z}_j}} \right). \tag{4.16}$$

The log profile likelihood (4.16) is exactly the log of Cox's partial likelihood with a time dependent categorical covariate, where the categories are defined by the time intervals $(x_{r-1}, x_r], r = 1, \ldots, R$. Instead of categorical covariate, we may also specify the values of x_r or any function of x_r at the event times.

A similar connection can be derived by letting the age be fixed and period interval lengths approach zero. The result will be the Cox proportional hazards model with (age) entering as time dependent strata or as a time dependent covariate in the model with basic time scale calendar time.

Other pairs of time scales are of course possible. For instance, dual time scales age and time since diagnosis arise frequently in clinical studies, age and time since weaning is another example from childhood life studies.

4.4 Simulation studies

4.4.1 Erroneous scale

The first simulation study investigates the performance of procedures to select relevant time scales discussed in Section 4.2. The procedures are the proportional hazards assumption test and frailty model estimation.

Several data generating models are assumed. We consider two competing time scales $\mathcal{S}_1$ with duration T_1 and $\mathcal{S}_2$ with duration T_2. $\mathcal{S}_1$ was specified as a better time scale than $\mathcal{S}_2$ in the sense that $\mathcal{S}_1$ has lower value of time delay δ_i than $\mathcal{S}_2$ has. One example in a real study, the true time scale is time since the onset of certain disease, T_1 is time since the subject feels any symptoms of the disease and T_2 is time since diagnosis. We assume that the time since onset can not be determined by the diagnosis.

The true duration T is generated by the ordinary proportional hazards model,

$$\lambda_i(t) = \lambda_0(t)\exp(\beta Z_i), \quad t > 0, \tag{4.17}$$

but we can only observe T_1 and T_2 generated from the true time scale with delays δ_i for each individual i. The details of the simulation procedure is described in the Appendix A-1. No truncation or censoring is considered in this simulation. Similar simulation studies have been considered by Liestøl and Andersen (2002) for the Gompertz-Makeham baseline hazard function with the purpose of showing the effect of misalignment patients and measurement error on the estimated regression coefficients.

To make $\mathcal{S}_1$ better than $\mathcal{S}_2$, the mean of δ_i for T_1 was specified lower than that for T_2 and δ_i follows uniform and exponential distributions. For this simulation the baseline hazards were determined parametrically as Gompertz, Exponential or Weibull hazard functions. One fixed categorical zero-one covariate Z_i generated the from Bernoulli distribution was also included.

Now, we compare the performance of the proportional hazard test (*ph*-test) and frailty variance estimation to detect the relevant time scales. The relevant time scale is expected to satisfy the proportional hazards assumption, and therefore will have larger p-values. In the frailty model, the estimated gamma frailty variance is used to detect the relevant time scale. A smaller frailty variance will indicate a better time scale. In a real situation, a more careful analysis can be performed. For example, a Schoenfeld residuals plot may be used to accompany the *ph*-test, and a confidence interval constructed from the profile likelihood may be calculated for the gamma frailty variance.

In the simulation the mean and standard deviation of the *ph*-test p-value and gamma frailty variance are used to summarize the result from 1000 replications. Histograms of the values are also examined (results not shown). In Tables 4.1 and 4.2 the mean of the *ph*-test p-value is under the *zph* column, and the mean of the gamma frailty variance is in the ω column.

There are some general comments for the generated data. The delays (δ_i) that follows an exponential distribution seems to make the model suffering from the violation of the proportional hazards assumption, shown by the low value of the coverage (the percentage

Table 4.1: Simulation study for erroneous scale with δ_i follows uniform distribution $U(0,1)$ and $U(0.5,2)$, for $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. The true coefficient β is 2. Each value is calculated based on a sample of size 200 with 1000 replications

Baseline	Model	Time Scale							
		$\mathcal{S}_1$				$\mathcal{S}_2$			
hazards		$\bar{\beta}$	p	zph	ω	$\bar{\beta}$	p	zph	ω
Gompertz	CPH	1.98	94.8	0.65	–	1.92	91.8	0.63	–
		0.19	–	0.25	–	0.19	–	0.25	–
	CPHF	1.99	94.2	0.66	0.009	1.99	94.2	0.66	0.009
		0.21	–	0.23	0.039	0.21	–	0.23	0.039
Exponential	CPH	1.64	47.8	0.44	–	1.43	9.6	0.41	–
		0.18	–	0.27	–	0.16	–	0.28	–
	CPHF	1.64	48.2	0.44	0.001	1.43	9.6	0.41	0
		0.18	–	0.27	0.01	0.16	–	0.28	0.001
Weibull	CPH	1.85	85.8	0.58	–	1.70	59.6	0.53	–
		0.19	–	0.26	–	0.19	–	0.27	–
	CPHF	1.85	85.8	0.59	0.004	1.70	60.0	0.54	0.002
		0.19	–	0.26	0.028	0.19	–	0.27	0.016

p is the coverage (percentage) of the interval estimation
$\bar{\beta}$ is the mean of estimated coefficient
zph is the mean of proportional hazards test p-value
ω is the mean of estimated frailty variance
The values in every second row are standard deviations

CoxPH : Cox's proportional hazards
CoxPHF : Cox's proportional
with frailty

Table 4.2: Simulation study for erroneous scale with δ_i follows exponential distribution with mean 0.5 and 1.25, for $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. The true coefficient β is 2. Each value is calculated based on a sample of size 200 with 1000 replications

Baseline	Model	Time Scale							
		$\mathcal{S}_1$				$\mathcal{S}_2$			
hazards		$\bar{\beta}$	p	zph	ω	$\bar{\beta}$	p	zph	ω
Gompertz	CPH	1.87	87.0	0.63	–	1.44	13.6	0.34	–
		0.20	–	0.25	–	0.20	–	0.30	–
	CPHF	1.90	87.0	0.66	0.02	1.65	48.8	0.57	0.169
		0.22	–	0.22	0.067	0.28	–	0.19	0.19
Exponential	CPH	1.23	1.2	0.36	–	0.70	0.0	0.13	–
		0.19	–	0.29	–	0.18	–	0.21	–
	CPHF	1.39	19.6	0.50	0.147	1.04	10.0	0.25	0.425
		0.26	–	0.21	0.196	0.36	–	0.18	0.417
Weibull	CPH	1.55	28.2	0.51	–	0.94	0.0	0.13	–
		0.18	–	0.29	–	0.18	–	0.20	–
	CPHF	1.63	46.2	0.61	0.069	1.31	22.6	0.34	0.375
		0.23	–	0.20	0.124	0.32	–	0.16	0.304

p is the coverage (percentage) of the interval estimation
$\bar{\beta}$ is the mean of estimated coefficient
zph is the mean of proportional hazards test p-value
ω is the mean of estimated frailty variance
The values in every second row are standard deviations

CoxPH : Cox's proportional hazards
CoxPHF : Cox's proportional
with frailty

of confidence intervals covering the true coefficient β). The most suffering one is the model with exponential baseline hazard.

When the baseline hazard function follows a Gompertz model, both the *ph*-test and frailty model show good performances. In Table 4.1, $\mathcal{S}_1$ and $\mathcal{S}_2$ are equally good, whereas in Table 4.2, $\mathcal{S}_1$ is better than $\mathcal{S}_2$ showed by the larger value of *zph* and smaller ω. The performances are confirmed by the coverage percentages p which have lower value for the wrong time scale.

Exponential baseline hazards are very much affected by the erroneous scale. Although the *zph*'s do not show very low values and ω's do not show very large values, the coverage probabilities are very low. In Table 4.1, it is rather hard to distinguish the time scales, because the values of *zph* and ω look similar, but the coverage probabilities are quite low for $\mathcal{S}_2$. For a larger effect of erroneous scale in Table 4.2, $\mathcal{S}_1$ and $\mathcal{S}_2$ can be distinguished by the value of *zph* and ω. The estimated frailty variances in the exponential baseline hazards (Table 4.1) do not seem to reveal the frailty effect. They give small variances but actually the effect is rather bad (lower coverages).

The performance of the procedures under the Weibull baseline hazard is generally similar with that of Gompertz. In the Weibull hazard the delays has a larger effect than in the Gompertz. The *zph* and ω can distinguish $\mathcal{S}_1$ and $\mathcal{S}_2$ in the data with a larger effect of erroneous scale.

When the procedures do not show a difference between $\mathcal{S}_1$ and $\mathcal{S}_2$, dual time scales modeling may be performed. For the data generated from these erroneous scale models, the inclusion of other time scales in the analysis will not likely increase the model fit.

4.4.2 Dual time scales

The second simulation study considered the approach discussed in Section 4.3 for modeling dual time scales $\mathcal{S}_1$ and $\mathcal{S}_2$. In a real study,

$\mathcal{S}_1$ and $\mathcal{S}_2$ could be calendar time and age, or age and time since diagnosis. In this simulation we assume the true model that generates duration T_1 as a result of using $\mathcal{S}_1$ follows a Cox model with time dependent covariate

$$\lambda_i(t \mid Z(t)) = \theta \exp(\beta_1 \eta_i + \beta_2 (t + \delta_i)), \quad t > 0. \tag{4.18}$$

For example, T_1 is time since onset of certain disease, T_2 is the age and δ_i is the age at onset, so where $T_1 = T_2 - \delta$. For positive β_2 the hazard for individual i will increase with time and the hazard is higher for individuals with higher δ_i (higher age at onset). The details of the data generating procedure of this simulation are presented in Appendix A-2.

In reality, we do not know the exact data generating process, we only believe that T_1 and T_2 should be modeled simultaneously. The compared performances for this simulation are the estimation of β_1 (the mean estimation, standard deviation) and the mean of the *ph*-test *p*-value (for analysis with Cox's model). One example of dual analysis is in the childhood mortality studies (Section 3.2). We may believe that the mortality depends on age and seasonality, therefore both time scales, age and period (as the proxy of seasonality), have to be taken care of. The variables of interest are not the times themselves but other explanatory variables such as gender, maternal education, etc. How the method of taking care of T_1 and T_2 affects the explanatory variable of interest is what we want to compare.

For the piecewise constant hazard approach, each time scale were divided into four equal-width intervals. Experimenting with several variations of gridding for generated data used in this simulation, four intervals gave reasonable piecewise constant hazards and was computationally feasible.

For the time-dependent strata, the same intervals as in the piecewise constant hazards were used. The analysis used the counting process data setup (Section 2.2). For each generated data set,

Table 4.3: Simulation study for dual time scales $\mathcal{S}_1$ and $\mathcal{S}_2$ analyzed with piecewise constant hazards and time dependent approaches. The true coefficients are $\beta_1 = 1.5$, $\beta_2 = 0, 1$ and δ_i is exponential with rate 0.85. Each value is calculated based on a sample of size 200 with 1000 replications

Method	$\beta_2 = 0$			$\beta_2 = 1$		
	$\bar{\beta}_1$ (sd)	p_1	zph	$\bar{\beta}_1$ (sd)	p_1	zph
piecewise const-hzd	1.55(.18)	95.2	–	1.51(.15)	95.5	–
$\mathcal{S}_1$ time-dep strata	1.51(.19)	94.9	.81	1.50(.18)	95.7	.71
$\mathcal{S}_2$ time-dep strata	1.51(.19)	95.2	.63	1.51(.18)	96.4	.66
$\mathcal{S}_1$ time-dep covariate	1.51(.18)	95.3	.90	1.48(.16)	95.7	.98
$\mathcal{S}_2$ time-dep covariate	0.91(.13)	5.6	.89	0.58(.13)	0.1	.99

$\mathcal{S}_1$ and $\mathcal{S}_2$ in front of the method's name denotes the basic time scale used

two time-dependent strata estimation procedures were carried out. The first one used $\mathcal{S}_1$ as the basic time scale with $\mathcal{S}_2$ as the time-dependent stratum, and the second used $\mathcal{S}_2$ as the basic time scale with $\mathcal{S}_1$ as the time-dependent stratum.

In the Cox time-dependent covariate analysis, the values of the covariate are only used at event times with a certain functional form. Analyzing time-dependent covariates in that way is computationally demanding, therefore we used similar time intervals as in the piecewise constant hazard and the time-dependent strata method. The form of the function is modeled non-parametrically using penalized smoothing spline (Hastie and Tibshirani, 1990), which is available, for instance, in the survival package of R or S-PLUS. As in the time-dependent strata case, two analyses were carried out by this model using each time scale and including the other time scale as a time-dependent covariate.

The results are shown in Table 4.3 for exponentially distributed δ_i and in Table 4.4 for uniformly distributed δ_i.

Table 4.4: Simulation study for dual time scales $\mathcal{S}_1$ and $\mathcal{S}_2$ analyzed with piecewise constant hazards and time dependent approaches. The true coefficients are $\beta_1 = 1.5$, $\beta_2 = 0, 1$ and δ_i is uniform(0,2). Each value is calculated based on a sample of size 200 with 1000 replications

Method	$\beta_2 = 0$			$\beta_2 = 1$		
	$\bar{\beta}_1$ (sd)	$\tilde{p}_1$	zph	$\bar{\beta}_1$ (sd)	$\tilde{p}_1$	zph
piecewise const-hzd	1.53(.17)	95.1	–	1.50(.15)	96.4	–
$\mathcal{S}_1$ time-dep strata	1.51(.18)	95.4	.81	1.49(.18)	95.7	.74
$\mathcal{S}_2$ time-dep strata	1.51(.18)	94.8	.61	1.50(.15)	97.6	.47
$\mathcal{S}_1$ time-dep covariate	1.50(.18)	95.1	.97	1.48(.17)	95.2	.95
$\mathcal{S}_2$ time-dep covariate	0.88(.13)	3.1	.92	0.50(.11)	0.0	.93

$\mathcal{S}_1$ and $\mathcal{S}_2$ in front of the method's name denotes the basic time scale used

In general, piecewise constant hazard and time-dependent strata show good performances. For the time-dependent strata, the appropriate analysis assuming model 4.18 is to use $\mathcal{S}_1$ as the basic time scale which gave good performances. However, even if the inappropriate basic time scale $\mathcal{S}_2$ is used, the performances are also good with only slightly violated proportional hazards assumption.

For the time-dependent covariate approaches, using $\mathcal{S}_1$ as the basic time scale gave good performances which is not surprising given the data generating model. Using the wrong basic time scale $\mathcal{S}_2$ is really harmful and worse if the time-dependent covariate is really in the model, i.e., $\beta_2 = 1$. Simulation with $\beta_2 = 0$ complements the result given by Liestøl and Andersen (2002). In their simulation T_1 is 'time since diagnosis' which had a Gompertz form and T_2 is age. The Gompertz baseline hazard is convenient since T_1 can switch into time-dependent covariate and still give the same result. In this simulation it is shown that a baseline hazard other than the Gompertz (constant hazard in this simulation) will give different results.

This issue has also been discussed for the case of an epidemiological follow-up study by Korn, Graubard and Midthune (1997).

4.4.3 Miss-specification

We also analyzed the generated data sets under miss-specified analysis, i.e., (i) the data was generated from the erroneous scale model but analyzed with the dual time scales methods; (ii) the data was generated from the dual time scales model but analyzed with the erroneous scale methods.

For the first miss-specified analysis, all dual time scales approaches showed good performances for the low effect of erroneous scale (the Gompertz baseline hazard case) but not for the large effect of erroneous scale (the exponential baseline hazard case). The Cox models with time-dependent strata and piecewise constant hazards approaches have similar performances and they are better than the Cox model with a time-dependent covariate.

For the second miss-specified analysis, the exponentially distributed δ_i, the *ph*-test and frailty model suggest that $\mathcal{S}_1$ is the most relevant time scale. However, for the uniformly distributed δ_i the procedures do not show any difference.

4.5 Application to infant mortality age–period analysis

We look again at the application considered in Section 3.2 about infant and child mortality. We mainly concentrate on the dual time scales age-period problem with categorical covariates *gender* (boy or girl) and *maternal education* (none, 6, 9, 12 years of education) for infant mortality data.

Analyses with piecewise constant hazards, Cox's proportional hazards with age time scale and Cox's proportional hazards with period time scale were performed. Two-month grids were applied for both age and period. For the piecewise constant hazards model, the standard Poisson model for the number of events in each grid with log link function was used. The total exposure times in each grid was entered to the model as an *offset*. For the Cox model with age time scale, period time was included as time-dependent strata or time-dependent covariate. Similarly, in the Cox model with period time scale, age was included as time-dependent strata or time-dependent covariate. The time-dependent covariates in both models using age and period as the basic time scale were treated non-parametrically using a penalized smoothing spline.

There is no scientific background suggesting that the two time scales, age and calendar time, are two alternative time scales. However, we can examine this by checking the proportionality assumption of the model using age and period as the basic time scale in separate analyses. No model violates the proportionality assumption with relatively large p-values for the proportionality test of 0.332 and 0.763 for age and period time scale, respectively.

Tables 4.5 and 4.6 show the result of likelihood ratio tests for the variables in each model and estimated coefficients, respectively. In this particular data set, in fact, they gave similar results. However, the safe approach is to consider the results from a Cox model with the time-dependent strata or piecewise constant hazards method. The general conclusion is that maternal education is quite important and gives protective effect in the case of infant mortality.

Table 4.5: Likelihood ratio test (LRT) for variables in the infant mortality models using piecewise constant hazards (pc-hazards), Cox proportional hazards with age time scale (Cox-age), Cox proportional hazards with period time scale (Cox-period)

Variables	pc-hazards	Cox-age		Cox-period	
		td-strata	td-covar	td-strata	td-covar
Age	$< .001$	—	—	—	$< .001$
Period	.395	—	.979	—	—
Gender	.080	.100	.075	.104	.132
Maternal educ.	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$

Table 4.6: Estimated coefficients and their standard errors for gender and maternal education in the infant mortality models

Variables	pc-hazards	Cox-age		Cox-period	
		td-strata	td-covar	td-strata	td-covar
Gender					
boy	—	—	—	—	—
girl	-0.31 (.18)	-0.29 (.18)	-0.32 (.18)	-0.29 (0.18)	-0.29 (.19)
Maternal educ.					
none	—	—	—	—	—
6 years	-0.76 (.26)	-0.76 (.27)	-0.74 (.26)	-0.72 (.27)	-0.53 (.28)
9 years	-1.17 (.34)	-1.18 (.35)	-1.16 (.34)	-1.11 (.35)	-1.10 (.36)
12 years	-2.74 (.56)	-2.72 (.56)	-2.70 (.56)	-2.67 (.56)	-2.43 (.57)

4.6 Remarks

The first consideration when we face a multiple time scales problem in event history analysis is to look for the scientific background of the time scales. The background may be obvious in clinical studies but may not be so in epidemiological or observational studies.

A proportional hazards test is advisable for checking the alternative time scales. This procedure is simpler than using a frailty model, moreover analyzing individual frailty may give wrong conclusions, especially when we use an incorrect underlying frailty distribution. We have noticed this problem also in the simulation studies.

A safe approach in analyzing dual time scales is to use the Cox model with time-dependent strata or the piecewise constant hazard approach. Simulation studies showed that both approaches gave good performance when analyzing dual time scales generated by the erroneous scale model or by the dual time scales model. The Cox model with time-dependent covariate is superior to other approaches when the other time scale (than the basic time scale) is really a time-dependent covariate in the model.

Chapter 5

Event History Analysis with Longitudinal Measurements

5.1 Introduction

We consider modeling event history with longitudinal measurements when the longitudinal measurements are intermittently observed and eventually measured with errors. Analysis of respiratory infection and weight in the ZINAK study presented in Chapter 3 is one example of such a situation.

One way of analyzing such data is by considering the time-to-event data as the outcome and longitudinal measurements as a time-dependent covariate. Another way is to analyze both outcomes simultaneously assuming that they are independent given certain latent processes.

Several methods have been proposed to deal with this kind of problem and some of them have been reviewed in Section 2.4.1. Four methods that have been around in the literature are LVCF (Last Value Carried Forward), TEL (time elapsed since the last measurement), two-stage, and joint model of event time and longitudinal measurements. Two methods based on Cox's proportional hazards model with stratification and frailty are proposed. The emphasis of the analysis is on the joint evolution of time-to-event and longitudinal measurements rather than longitudinal measurements as surrogate markers for the event.

To our knowledge, all methods mentioned above were mostly applied to clinical settings such as AIDS studies, psychiatric disorders and cancer prevention trials, not to observational or epidemiological settings which are more ”irregular” than the clinical ones. Applications of the methods to multiple events or repeated events are also rarely considered in the literature.

The aim of this chapter is to compare the methods by means of simulation and to perform further analysis of the respiratory infection and weight data from the ZINAK study introduced in Chapter 3.

5.2 Problem and models

Suppose n individuals are followed over a time interval $[0, L)$ with *longitudinal measurements* $\{y_{ij} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, n_i\}$ at times $\{t_{ij} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, n_i\}$. Together with the measurements, a *counting process* $\{N_i(u) : 0 \leq u \leq L\}$ for the events and a predictable *at risk* process $\{K_i(u) : 0 \leq u \leq L\}$ are also recorded. An additional fixed time covariate or baseline covariate $\mathbf{Z}$ may be included.

One example of such a data setup is illustrated in Figure 5.1. The event history data are repeated events data in which one individual may have several counting process intervals `(`t_0`,` t_1`], event`. The at-risk process $\{K_i(u) : 0 \leq u \leq L\}$ is alternating between 0 and 1 at time points specified by the intervals. Notice that for repeated events such as morbidity, after an event occurrence, the individual is not at risk for a certain period of time. The not-at-risk period corresponds to the duration of illness (denoted by dashed lines in Figure 5.1(a) under event-symptoms).

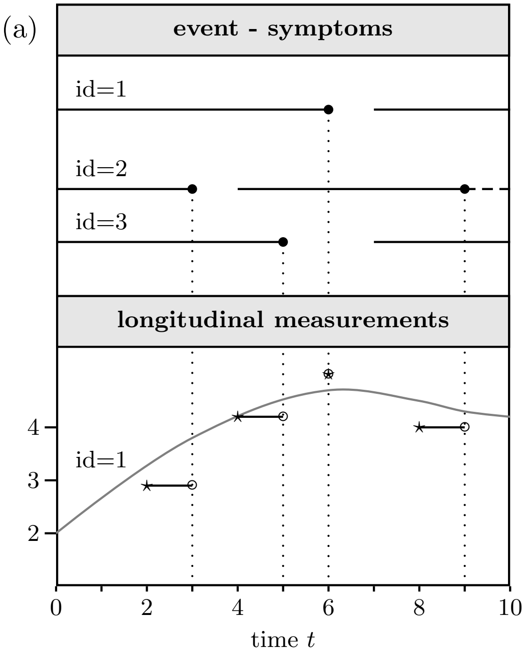
The longitudinal measurements are obtained throughout the period of observation and do not necessarily coincide with the event times. The observed measurement data are not perfect since the true time-dependent covariate might be a continuous curve as depicted in the figure but we only collect some values (the $\star$'s in the picture, for id = 1). Moreover, the values may be subject to measurement errors (the $\star$'s are not exactly on the curve). This is a quite common situation in many applications, for instance when measuring infant weights. The situation creates a problem when we use Cox's model for analyzing event history data since the partial likelihood requires the values of all covariates at the event times (see Equation (2.30)).

We consider two models for the data generating mechanism of time-to-event and longitudinal outcomes. The first one assumes a model of time-to-event with the longitudinal measurements as a time-dependent covariate. The second one assumes a *joint model* of time-to-event and longitudinal measurements induced by a latent process.

In general, any model may be specified for the longitudinal measurements. Here, we consider a linear random effects model for the longitudinal measurements as in the infants' growth model of the ZINAK study (Section 3.5). The covariate process for the infant weight data may be specified as

$$\mathbf{Y}^\star(t) = \alpha_0 + \boldsymbol{\alpha}_1\mathbf{Z} + \alpha_2 t + \mathbf{U}_1 + \mathbf{U}_2 t, \quad t > 0, \tag{5.1}$$

where $\mathbf{Y}^\star(t) = (Y_1^\star(t), \ldots, Y_n^\star(t))$ is the "true" weight at age t for individual $i = 1, \ldots, n$; $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ are fixed time covariates;



(b)

id	t0	t1	event
1	0	6	1
1	7	10	0
2	0	3	1
2	4	9	1
3	0	5	1
3	7	10	0

(c)

id	t	Y
1	2	2.9
1	4	4.2
1	6	5
1	8	4

Figure 5.1: (a) Event history data and longitudinal measurements and an illustration of imputing time-dependent covariate values using LVCF and time since measurements; (b) event history data (c) longitudinal data

$\mathbf{U}_1$ and $\mathbf{U}_2$ are the unobservable random effects for the intercept and age t with $(\mathbf{U}_1, \mathbf{U}_2) \sim N(\mathbf{0}, \mathbf{\Sigma})$. The observed covariate process is

$$\mathbf{Y}(t) = \mathbf{Y}^\star(t) + \epsilon_t, \quad t = t_1, t_2, \ldots, \tag{5.2}$$

where $\epsilon_t \sim N(0, \sigma)$ are the measurements errors. The actual observation $\mathbf{Y}(t)$ is not continuously observed but finitely observed at times $t_1, t_2, \ldots$.

The time-to-event is modeled through Cox's proportional hazards model

$$\lambda(t \mid \mathbf{X}, Y^\star(t)) = \lambda_0(t) \exp(\boldsymbol{\beta}_1 \mathbf{X} + \beta_2 Y^\star(t)), \tag{5.3}$$

where $\mathbf{X}$ are fixed time covariates such as *maternal education*, *supplementation* and may also include $\mathbf{Z}$ (e.g., the *gender* variable in the covariate process model).

The central methodological and practical problem is how to estimate the parameters in (5.3) when $Y^\star(t)$ is not available but $Y(t)$ is instead. The methods we consider are LVCF, TEL, two-stage, Cox-frailty and Cox-strata discussed in the next section.

The joint model was mentioned in Section 2.4.2 as a more general methodology to deal with time-dependent covariates. It has two submodels, one for the longitudinal measurements and another for the time-to-event.

The longitudinal measurements model is the same as model (5.2). The specification of the time-to-event model, however, is different from that of (5.3). The hazard function of this joint model is

$$\lambda(t \mid \mathbf{X}) = \lambda_0(t) \exp(\boldsymbol{\beta} \mathbf{X} + \gamma(\mathbf{U}_1 + \mathbf{U}_2 t)), \tag{5.4}$$

where $\mathbf{X}$ are fixed time covariates, could be the same or overlap with $\mathbf{Z}$; $\mathbf{U}_1$ and $\mathbf{U}_2$ are specified similarly as in model (5.3). The difference compared to model (5.3) is that the "true" value $Y^\star(t)$ are not included in the hazard function but only the random effect part $\mathbf{U}_1 + \mathbf{U}_2 t$.

The idea of the joint model is that the dependence between the longitudinal measurements and the time-to-event can arise through the common covariate $\mathbf{X}$ and the possible unobserved heterogeneity in both models. The joint model attempts to take care of the latent heterogeneity in both models simultaneously. When there is no latent association, $\gamma = 0$, the joint model is actually two separate models of longitudinal data and event history data.

It is also possible to include more random effect terms than $\mathbf{U}_1 + \mathbf{U}_2 t$. The latent association can also be extended as in the Henderson et al.'s (2000) model and there can even be more than one longitudinal measurement (Lin, McCulloch and Mayne, 2002). However, in many practical situations, the random effects term of the initial value of measurement (the intercept) and the steepness of the longitudinal covariate by time (e.g., age) are the most important terms.

5.3 Methods

The four methods of LVCF, TEL, two-stage and joint model have been reviewed briefly in Section 2.4.2. We discuss the methods further here with some illustrations.

Suppose we have event history data and longitudinal data as in Figure 5.1. To construct LVCF and TEL for the individual with id = 1, we have to know the covariate value for this individual at the event times 3, 5, 6, and 9.

The values obtained by the LVCF method are the symbol $\circ$ in Figure 5.1. They are obtained by assuming that the most recent measurement value is the value at event-time. It is possible that the event-times correspond to the covariate-time (as at $t = 6$, the symbol ⍟). The time elapsed since the last measurement (TEL) is the length of the horizontal line connecting the actual measurement and the event time.

The LVCF and TEL methods are used in the ordinary Cox regression as

$$\lambda(t \mid \mathbf{Z}, \tilde{Y}_t) = \lambda_0(t) \exp(\beta \mathbf{Z} + \tilde{Y}_t), \tag{5.5}$$

$$\lambda(t \mid \mathbf{Z}, \tilde{Y}_t, \tau) = \lambda_0(t) \exp(\beta \mathbf{Z} + \tilde{Y}_t f_1(\tau_t) + f_2(\tau_t)), \tag{5.6}$$

respectively, where $\mathbf{Z}$ are fixed-time covariates; $\tilde{Y}_t$ denotes the value obtained by LVCF, τ_t is the TEL at t, f_1 and f_2 are suitable functions.

The LVCF method is known to give biased estimates of the parameters (Prentice, 1982). However, we believe that LVCF is commonly used in practice. The $\tilde{Y}_t$ could be a good predictor of hazard when the effect of the longitudinal measurements on event time is delayed.

The idea of the TEL method is that the measurements could be "aging" and new information closer to the event time is better than old information. When the measurements are irregularly observed, the value of τ_t (TEL) could carry information about the subjects disease progression (Bruijne et al., 2001). However, it is not likely to be an added advantage for regular measurements such as in the ZINAK study. The added difficulty in using TEL instead of LVCF is in specifying f_1 and f_2. Therefore, the skill and tool needed in TEL is actually the same as modeling the ordinary Cox regression. This method could be a practical alternative to more complex sophisticated methods.

The two-stage method is mentioned briefly in Section 2.4.2. The main idea of the method is to reconstruct the covariate function given the observed values of the covariate. In the two-stage approach, Cox's proportional hazards model becomes

$$\lambda(t \mid \mathbf{Z}, Y_t) = \lambda_0(t) \exp(\beta_1 \mathbf{Z}) E\left[\exp(\beta_2 Y^\star(t)) \mid K(t) = 1, Y_t, \mathbf{Z}\right], \tag{5.7}$$

where $\mathbf{Z}$ are fixed-time covariates always available at the event times; Y_t denotes the observed values (the $\mathbf{Y}(t)$ in model (5.2)); and $K(t) = I_{\{T \geq t\}}$ is an at-risk process indicator function (the usual notation $Y(t)$ for the at-risk process has been reserved for the longitudinal measurements). Tsiatis et al. (1995) used a first-order approximation of the conditional expectation in model (5.7).

The LVCF, TEL and two-stage methods basically assume the first model discussed in Section 5.2 (Equations (5.1), (5.2), and (5.3)) where the central problem is on imputing missing longitudinal covariate values at the event times. The two-stage model is more computationally demanding than the others since it needs to fit a longitudinal model at each event time.

The joint model method assumes the joint model of the time-to-event and longitudinal measurements process induced by a latent process discussed in Section 5.2. We have mentioned several methods to fit the model in Section 2.4.2. Basically, the joint method maximizes the likelihood function of longitudinal and hazard model simultaneously. Theoretically, it has many desirable properties, such as less biased parameter estimates, making the efficient use of data and easier model validation (Tsiatis et al., 1995; Henderson et al., 2000; Ibrahim, Chen and Sinha, 2001; Tsiatis and Davidian, 2004). Practically, the methods developed for this model are still lacking computational tools and this model is computationally demanding (Do, 2002).

In addition to the methods discussed above, we propose two methods based on Cox's model with stratification and Cox's model with frailty. We call them *Cox-strata* and *Cox-frailty*, henceforth. The main idea of both methods is adjustment for the longitudinal covariate when the covariate is considered as a nuisance variable. For example, when the interest is not on the effect of weight on the morbidity but on the effect of other variables such as *supplementation*, *gender* or *maternal education*, these methods could be a reasonable choice.

Basically we assume a constant multiplicative effect of $\exp(Y^\star(t))$ on the baseline hazard function over time. This assumption may be violated when we use $Y(t)$ as a proxy of $Y^\star(t)$. Stratification is the usual approach to deal with non-proportionality.

In this longitudinal measurements problem, the $Y(t)$ is time dependent, therefore the stratification is actually a time-dependent stratification. Cox's model with time-dependent strata is

$$\lambda(t \mid \mathbf{Z}, Y(t)) = \lambda_{0y_j}(t) \exp(\beta\mathbf{Z}), \quad \text{if } Y(t) = y_j, \tag{5.8}$$

where y_j is the value of $Y(t)$, $j = 1, 2, \ldots, V$, where V is the number of unique values of $Y(t)$.

In practice, when V is large, the Cox-strata method may not be feasible and the precision of estimated coefficients may be low. To overcome this, we may categorize $Y(t)$ such that the size of V is reasonable.

The term $\exp(\beta_2 Y^\star(t))$ in model (5.3) which is a random variable instead of a fixed variable also leads naturally to Cox's model with frailty (Section 2.2.5, Equation (2.13)). In this case, clusters in this frailty model are the longitudinal measurements, therefore we use the value of the longitudinal measurements as a categorical variable, the same as y_j in the Cox-strata method. In fact, this frailty approach is one alternative solution when the V in the Cox-strata method is large.

Practically, we may use the value obtained by the LVCF method or by the two-stage method for the value of $Y(t)$. The problem of specifying the distribution of the frailty effects is the same as in any Cox's frailty model. We have discussed this problem for the childhood mortality model (Chapter 3) and the multiple time scales problem (Chapter 4).

5.4 Simulation studies

The purpose of the simulation study is to investigate the performance of the methods discussed in the previous section. We compared the LVCF, TEL, two-stage, Cox-frailty and Cox-strata methods. The joint model was not included in the comparison since, as we have noticed in the previous section, it requires heavy computation and is not feasible for large simulations.

The simulated data for each individual consists of event history data `(`t_0, t_1`]`, `event` with one fixed covariate Z and longitudinal data (t, Y_t), similar to the illustration in Figure 5.1. The fixed covariate could be *supplementation*, *gender* or *maternal education* as in the ZINAK study. The details of the simulation procedures are found in Appendix A-3. Simulations were performed in ***R*** (Ihaka and Gentleman, 1996; R Development Core Team, 2004).

We look at $\hat{\beta}_1$, the estimated coefficient of Z, as one criterion of method performances. The coverage, an indicator whether the interval estimation includes the true parameter or not, were calculated. Proportionality of the hazard model is another criterion to be investigated. Additionally we also looked at $\hat{\beta}_2$, the estimated coefficients of the longitudinal measurements Y, for the LVCF, TEL and two-stage methods.

Tables 5.1 and 5.2 show the results of the simulations based on the two main models discussed in Section 5.2. When there is no covariate effect in the hazard model (5.3) or no latent association in model (5.4), that is $\beta_2 = 0$, all methods arrived at similar results. The performance of the methods in estimating β_1 were similarly good with coverages close to 95%. The performance investigated from the proportionality assumption was also good for all methods but the TEL method generally has better hazard proportionality than the others.

For the LVCF, TEL and two-stage methods, we can also look at

Table 5.1: Simulation study for Cox's time-dependent covariate model analyzed with the LVCF, TEL, two-stage, Cox-frailty and Cox-strata methods. See the text and Appendix A-3 for the simulation specifications. Each value is calculated based on a sample of size 50 with 500 replications

Method	$\beta_2 = 0$			$\beta_2 = -0.1$		
	$\bar{\beta_1}$ (sd)	p_1	*zph*	$\bar{\beta_1}$ (sd)	p_1	*zph*
LVCF	1.23 (.170)	94.3	.477	1.24 (.232)	95.2	.498
TEL	1.23 (.170)	94.1	.547	1.24 (.233)	95.2	.580
Two-stage	1.22 (.167)	95.5	.507	1.24 (.236)	95.3	.505
Cox-strata	1.23 (.271)	94.7	.510	1.24 (.361)	95.0	.492
Cox-frailty	1.23 (.168)	95.3	.492	1.24 (.232)	95.6	.490

$\bar{\beta_1}$ is the mean of the estimated coefficient $\hat{\beta_1}$ (true value $\beta_1 = 1.2$) and their standard deviation (sd) in parentheses

p is the coverage (percentage) of the interval estimation

zph is the mean of the proportional hazards test p-value

Table 5.2: Simulation study for joint model analyzed with the LVCF, TEL, two-stage, Cox-frailty and Cox-strata methods. See the text and Appendix A-3 for the simulation specifications. Each value is calculated based on a sample of size 50 with 500 replications

Method	$\beta_2 = 0$			$\beta_2 = 1$		
	$\bar{\beta_1}$ (sd)	p_1	*zph*	$\bar{\beta_1}$ (sd)	p_1	*zph*
LVCF	1.20 (.193)	94.8	.512	1.01 (.254)	78.4	.429
TEL	1.20 (.196)	94.6	.589	1.01 (.257)	78.4	.510
Two-stage	1.22 (.203)	94.5	.493	1.02 (.259)	78.6	.441
Cox-strata	1.22 (.365)	96.0	.515	1.03 (.447)	85.1	.478
Cox-frailty	1.20 (.160)	95.5	.508	1.04 (.207)	84.5	.476

$\bar{\beta_1}$ is the mean of the estimated coefficient $\hat{\beta_1}$ (true value $\beta_1 = 1.2$) and their standard deviation (sd) in parentheses

p is the coverage (percentage) of the interval estimation

zph is the mean of the proportional hazards test p-value the performance based on the estimates of β_2. The β_2 were almost perfectly estimated with mean of 0.003, 0.002, and 0.002; coverage 94.9%, 95.7% and 94.1%, for the LVCF, TEL and two-stage methods, respectively.

When data was generated by Cox's time-dependent covariate model and the effect of covariate Y was present, i.e., $\beta_2 = -0.1$ (Table 5.1), the performance of all methods were as good as that of the result with $\beta_2 = 0$. It is rather surprising that the LVCF method is similarly good enough to estimate β_1 for this situation as compared to the two-stage method, except perhaps for its proportionality. The LVCF method is probably good when the time-dependent covariate is a linear model with low gradient, as specified for Y in this simulation.

For the LVCF, TEL and two-stage methods, their estimates of β_2 were rather poor with means of -0.005, -0.013, and -0.005, respectively; coverage 82.2%, 89.8% and 83%, respectively. This problem was due to the measurement errors of Y. The estimation of β_2 was slightly better with the TEL method.

When data were generated from the joint model with latent association (Table 5.2), the estimates of β_1 were biased for all methods. Here, the Cox-strata and Cox-frailty methods are slightly better than the LVCF, TEL and two-stage methods with larger coverage probability. The performances dramatically went down for the LVCF, TEL and two-stage methods in estimating β_2. Their estimates were close to zero and far from the true value 1 of β_2. Their coverage were close to zero as well. These severe under-estimations were largely caused by the miss-specification of Y.

5.5 Application to infant respiratory infection and weight data

We continue the analysis of infants respiratory infection and weight data from the ZINAK study introduced in Chapter 3. We included the weight longitudinal covariate in the hazard model and analyzed the data using the LVCF, TEL, two-stage, Cox-strata and Cox-frailty, and joint model methods.

We used both the Andersen-Gill model (AG model) and the gap-time model to specify the model of repeated events. We expect improvements over the models presented in Table 3.7 (the AG model) and Table 3.8 (the gap-time model) by using these methods.

The implementation of the LVCF method to the data is straightforward but rather tedious. First we have to arrange the data by event-time splitting (See example in Figure 5.1, Section 5.2). There were 2,423 records for this repeated time to event data set collected from 666 subjects. The number of records grew considerably into 104,838 records after splitting. The missing values of weight at event times were then imputed by the values of weight from 3,770 records of the weight data set.

The τ (the time since the most recent measurements) in the TEL method can be directly calculated from the LVCF data set. There are many alternatives to model f_1 and f_2, parametrically or non-parametrically. We consider parametric models of the exponential form.

Cox and Oakes (1984); Bruijne et al. (2001) considered the exponential form

$$\beta_1 Y_{(t-\tau)} + \beta_2 Y_{(t-\tau)} \exp(\mathrm{C}\tau) + \beta_3 \exp(-\mathrm{D}\tau), \qquad (5.9)$$

where $Y_{(t-\tau)}$ is the measurements obtained by the LVCF method, and τ is the time since the most recent measurements, C and D are constant parameters which give the highest maximized log-likelihood of the model. We do not elaborate this form and chose the simplest C = D = 1. As we can see later, this simple form of the TEL model had the best fit compared to the LVCF and two-stage methods.

In the two-stage method, a linear random effects model was used for the weight growth curve model as in Equation (3.2) but only included *gender* as a fixed covariate. The two-stage method was the most time consuming in the data preparation since a new model had to be fitted at each event time.

The Cox-strata and Cox-frailty methods used the actual value of weight obtained by the LVCF method as the stratum or cluster in the models. The weight data were measured in 2 digits precision. For Cox-strata, the values were rounded into 1 digit, giving about 90 unique values of weight as strata.

The results from the LVCF, TEL and two-stage methods were actually quite similar, especially the LVCF and two-stage methods were close. The likelihood ratio test in Table 5.3 shows that the two-stage method was only slightly better than the LVCF method but the TEL method certainly had the best goodness of fit among the others, both for the Andersen-Gill and gap-time repeated events models. We only present the result from the TEL method in the subsequent discussion.

The results from the Cox-frailty method were almost identical to the previous AG and gap-time models (Table 3.7 and Table 3.8). The random effect variance was very small and its likelihood ratio test was not significant. The results from Cox-strata generally included wider confidence intervals than that from the other models. We only present the results from the Cox-strata together with the results from the TEL method.

Tables 5.4 and 5.5 present the results of the fitted model using the TEL and Cox-strata methods from the previous models in Tables

Table 5.3: Likelihood ratio test for the LVCF, TEL and two-stage models compare to the model in Table 3.7 for the AG model and Table 3.8 for the gap-time model

Method	AG model		gap-time model	
	Deviance(df)	p-val	Deviance(df)	p-val
LVCF	2.9 (1)	.087	12.7(1)	.0004
TEL	13.4(3)	.004	22.0(3)	.00005
two-stage	3.1 (1)	.076	13.1(1)	.0002

3.7 and 3.8. Note that the reference categories in the variables were omitted for conciseness.

The estimates of *gender*, *supplementation* and *maternal education* were similar to that of Table 3.7 (the AG model without weight variable and τ). Maternal education and gender had an important contribution to the model. The Cox-strata method was generally conservative with wider confidence intervals as compared to that of the TEL method and results in Table 3.7. The proportionality assumptions for these models were checked, there was no indication of a proportional hazard violation.

The weight variable was only slightly important in the model but the statistical interaction with $\exp(\tau)$ was very significant. Weight seemed to have a protective effect from respiratory infections in infants. Removing weight from the model and its interaction with τ did not improve the fitted model, therefore both weight, τ and their interaction were kept in the model. Similar results were found in the model using the gap-time repeated event model of Table 5.5.

Finally, we compared the above methods with the joint model method. The hazard model was specified as an exponential hazard model with all the variables as in Table 5.4 or 5.5 included, except the *weight* and *tel* variables. As we have found in Section 3.4 and

Table 5.4: Hazards model for respiratory infection using the Andersen-Gill model

Parameter	TEL method		Cox-strata method	
girl	0.91	(0.81, 0.98)	0.89	(0.80, 0.99)
zinc	1.01	(0.88, 1.15)	1.02	(0.89, 1.17)
zinc+iron	0.91	(0.80, 1.04)	0.93	(0.80, 1.06)
iron	0.96	(0.85, 1.10)	0.96	(0.83, 1.10)
6 years	0.85	(0.64, 1.11)	0.86	(0.64, 1.16)
9 years	0.71	(0.54, 0.94)	0.70	(0.52, 0.95)
12 years or more	0.48	(0.30, 0.77)	0.50	(0.30, 0.82)
weight	0.96	(0.91, 1.01)		
exp(tel)	0.99	(0.98, 1.00)		
weight*exp(tel)	1.002	(1.001, 1.003)		

tel is the time since the most recent measurement of weight
The estimated parameters are presented as $\exp(\hat{\boldsymbol{\beta}})$

Table 5.5: Hazards model for respiratory infection using the gap-time model

Parameter	TEL method		Cox-strata method	
girl	0.87	(0.79, 0.96)	0.87	(0.79, 0.96)
zinc	1.01	(0.88, 1.15)	0.99	(0.87, 1.13)
zinc+iron	0.91	(0.80, 1.04)	0.91	(0.79, 1.04)
iron	0.97	(0.85, 1.10)	0.95	(0.83, 1.08)
6 years	0.85	(0.65, 1.12)	0.82	(0.62, 1.08)
9 years	0.72	(0.54, 0.95)	0.69	(0.52, 0.91)
12 years or more	0.48	(0.30, 0.77)	0.46	(0.29, 0.75)
weight	0.92	(0.87, 0.96)		
exp(tel)	0.99	(0.99, 1.00)		
weight*exp(tel)	1.002	(1.001, 1.003)		

tel is the time since the most recent measurement of weight
The estimated parameters are presented as $\exp(\hat{\boldsymbol{\beta}})$

also in this section, the results for the AG model and the gap-time model were similar, indicating that an exponential baseline hazard should be fine. The longitudinal model was specified similarly as in Table 3.9 (random effect growth curve model). The estimate is based on the joint maximized likelihood of exponential hazard model and linear random effect model. ***SAS*** with the ***NLMIX*** procedure (Guo and Carlin, 2004) was used to fit the joint model. With 2,423 records of event history data and 3,172 records of longitudinal data, it took 35 hours to fit the model. The result is presented in Table 5.6.

The hazard model in the separate analysis column in Table 5.6 is comparable to that of Table 5.5 or Table 3.8 since the exponential model fitted in the joint model also used gap-time as the time scale. The longitudinal model in the separate analysis column is the same as Table 3.9.

Although small, the estimated risk ratios for the hazard model generally were away from one as compared to the separate model, indicating that the possible frailty effect in the model had been taken care of by the joint model. For the longitudinal model, the random effect components were stronger than in the separate model, which may indicate that possible under-estimations had been taken care of. The general conclusion for the effect of the variables, however, is similar to that of separate models.

The significant latent association γ gave additional information about the positive association between random effects from both models. In this type of joint model, we may interpret the γ as the effect of time-dependent frailty in the hazard model which operates through age. The positive value indicates that age had a considerably large effect on the hazard of experiencing respiratory infection. We compared this finding with an ordinary gap-time Cox-frailty model using age as a frailty term. The analysis was performed by adding the frailty term age in the TEL model of Table 5.5. We found that the variance of the random effect was 1.03 with a very significant result of LRT (497.6 with 1 degree of freedom) which therefore confirms the significant latent association in the joint model.

We summarize the findings for the infants' respiratory infection and weight data. Maternal education seemed to be important for infant respiratory infection but not for weight. Weight was associated with infant respiratory infection and its duration. Finally, none of the methods show any statistical significance of supplementation. However, we have not considered other important variables such as breastfeeding, food intake and socio-economic indicators in the model, further analyses with those variables may be necessary.

5.6 Remarks

Cox based models such as the LVCF, TEL, and two-stage methods should be good enough in situations where data comes a from Cox's model with time-dependent covariate. Practically, the TEL method would be the first choice. The TEL method may even be used when the measurement is only performed once, in this situation the two-stage or the joint model methods may be difficult to perform. The TEL method is also favorably applied to the *switching-treatment* type covariate where the covariate path is a step function with only a few values during the period of observation instead of continuous function covariate.

Care must be taken in using the Cox based model with a time-dependent covariate under the model with miss-specification. The Cox-strata or Cox-frailty may be more appropriate in the situation when the longitudinal covariate is regarded as a nuisance variable, in which there is no need to explicitly estimate their effects. Alternatively, the joint model method can be used, but this may require complex and heavy computation.

Table 5.6: Separate and joint model analysis for infant respiratory infection and weight data

Parameter	Separate analysis		Joint analysis	
		hazard model		
Intercept	0.64	(0.49, 0.85)	0.53	(0.35, 0.79)
girl	0.91	(0.83, 1.00)	0.92	(0.81, 1.05)
zinc	1.00	(0.88, 1.14)	0.98	(0.82, 1.17)
zinc+iron	0.91	(0.79, 1.03)	0.90	(0.75, 1.08)
iron	0.97	(0.85, 1.11)	0.98	(0.82, 1.17)
6 years	0.84	(0.64, 1.10)	0.86	(0.58, 1.28)
9 years	0.70	(0.53, 0.92)	0.72	(0.48, 1.08)
12 years or more	0.46	(0.28, 0.74)	0.47	(0.25, 0.88)
		longitudinal model		
Intercept	6.37	(5.88, 6.86)	6.37	(5.89, 6.84)
Age	0.17	(0.17, 0.18)	0.17	(0.16, 0.18)
girl	-0.54	(-0.68, -0.40)	-0.54	(-0.68, -0.40)
zinc	0.02	(-0.18, 0.22)	0.02	(-0.18, 0.21)
zinc+iron	0.01	(-0.19, 0.21)	0.01	(-0.18, 0.20)
iron	0.01	(-0.19, 0.21)	0.01	(-0.18, 0.20)
6 years	0.20	(-0.27, 0.68)	0.20	(-0.26, 0.67)
9 years	0.31	(-0.17, 0.79)	0.31	(-0.16, 0.78)
12 years or more	0.26	(-0.41, 0.94)	0.26	(-0.39, 0.92)
Illness days	-0.53	(-0.64, -0.41)	-0.53	(-0.65, -0.41)
Random effects				
sd(Intercept)	0.993	(0.923,1.064)	0.997	(0.927, 1.068)
sd(Age)	0.065	(0.061,0.070)	0.067	(0.062, 0.072)
corr(Intercept, Age)	-0.617	(-0.860,-0.430)	-0.684	(-0.940,-0.486)
		latent association		
γ	-	-	0.596	(0.500, 0.692)

For the hazard models, the estimated parameters are presented as $\exp(\hat{\boldsymbol{\beta}})$
γ is the parameter specified in Equation (5.4)

Chapter 6

Concluding Remarks

This thesis has contributed several solutions and discussions to the problems in event history analysis with multiple time scales and longitudinal measurements motivated by some epidemiological studies.

The focus is on the Cox regression model, but this is by no means the solution to all problem. Other approaches such as the parametric proportional hazards, additive hazards and accelerated failure times, that have been omitted in the discussion, deserve attention. Similar problems presented in this thesis will certainly appear in those approaches as well.

We have presented methods for choosing the time scale in the Cox regression model based on the proportional hazards test and the frailty model. Although the methods are inferential, we suggest using the methods as exploratory tools together with consideration of the scientific background of the data.

When several time scales are considered to be important and the model is a pure bivariate or multivariate time scale model, the Cox model with time-dependent strata, or the piecewise constant hazards approach are suggested. The price is that, in the Cox model with time-dependent strata the effect of the time scale can not be quantified by means of the estimated regression coefficients; and the piecewise constant hazards is only an approximation of the model. A general methodology for this multivariate time scale problem still needs more investigation. The developments since the review by Andersen et al. (1993, Chapter X) are the non-parametric estimation of the bivariate survivor function(Prentice, 1999; Gentleman and Vandal, 2002) and a more theoretical ground by Ivanoff and Merzbach (2002).

We have presented comparisons of several widely used methods to deal with longitudinal measurements in the event history analysis together with two proposed methods. Comparison by simulation showed that the time elapsed measurement time method (TEL) performed well when the data came from the Cox model with a time-dependent covariate. The two proposed methods based on Cox's model with stratification and frailty may be useful when the data are suspected to cause miss-specification in the Cox model. In the comparison by simulation we have left out the joint model, a promising method that unfortunately requires heavy and complex computation. The joint model is not in a mature development state yet, especially in the computing aspects. Further research is certainly needed. An estimation method in the generalized linear latent models (Huber, Ronchetti and Victoria-Feser, 2004) seems to be fruitful to estimate the joint model. Other urgent future research is diagnostic tools for the joint model, which is still in its infancy.

Finally, any developed methods should have a real advantage in practice. We have performed several analyses by the discussed methods using epidemiological surveillance and randomized trial data. We have confirmed the results obtained by the original investigators and contributed additional insights to their findings.

Bibliography

Aalen, O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics* **6**: 701–726.

Andersen, P. (2003). Two encyclopedia contributions: Time-dependent covariate, *Technical report*, Department of Biostatistics, Institute of Public Health, University of Copenhagen.

Andersen, P. K. (1991). Survival analysis 1982-1991: The second decade of the proportional hazards regression model, *Statistics in Medicine* **10**: 1931–1941.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag Inc.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis, *Statistical Methods in Medical Research* **11**(2): 91–115.

Andersen, P. K. and Liestøl, K. (2003). Attenuation caused by infrequently updated covariates in survival analysis, *Biostatistics* **4**: 633–649.

Bailey, K. R. (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model, *The Annals of Statistics* **12**: 730–736.

Bates, D. M. and Pinheiro, J. (1998). Computational methods for multilevel models., *Technical memorandum bl0112140-980226-01tm*, Bell Labs, Lucent Technologies, Murray Hill, NJ.

Berzuini, C. and Clayton, D. (1994a). Bayesian analysis of survival on multiple time scales, *Statistics in Medicine* **13**(8): 823–838.

Berzuini, C. and Clayton, D. (1994b). Bayesian analysis of survival on multiple time scales, *Statistics in Medicine* **13**: 823–838.

Bhandari, N., Bahl, R., Mazumdar, S., Martines, J., Black, R., Bhan, M. and Infant Feeding Study Group (2003). Effect of community-based promotion of exclusive breastfeeding on diarrhoeal illness and growth: A cluster randomised controlled trial, *Lancet* **361**: 1418–1423.

Black, R., Morris, S. and Bryce, J. (2003). Where and why are 10 million children dying every year?, *Lancet* **361**: 2226–2234.

Broström, G. (2002). Cox regression; ties without tears, *Communications in Statistics, Part A – Theory and Methods* **31**(2): 285–297.

Bruijne, M. H. J. d., Cessie, S. l., Kluin-Nelemans, H. C. and Houwelingen, H. C. v. (2001). On the use of Cox regression in the presence of an irregularly observed time-dependent covariate, *Statistics in Medicine* **20**(24): 3817–3829.

Central Bureau of Statistics (CBS) [Indonesia], State Ministry of Population/National Family Planning Coordinating Board (NFPCB) and Ministry of Health (MOH) and Macro Intemational Inc. (MI) (1998). *Indonesia Demographic and Health Survey 1997*, CBS and MI., Calverton, Maryland.

Clayton, D. (1988). The analysis of event history data: A review of progress and outstanding problems, *Statistics in Medicine* **7**: 819–841.

Commenges, D. (1999). Multi-state models in epidemiology, *Lifetime Data Analysis* **5**: 315–327.

Cox, D. R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **34**: 187–220.

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

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

Danardono (2000). *Multilevel Model of the Diarrhea Occurrence in Children*, Master's thesis, Department of Biostatistics and Demography, Faculty of Public Health Khon Kaen University, Thailand.

Danardono (2003). Event history analysis of childhood mortality and morbidity in Purworejo, Indonesia., *Statistical studies 30*, Department of Statistics, Umeå University.

Diggle, P. (1988). An approach to the analysis of repeated measurements, *Biometrics* **44**: 959–971.

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, second edn, Oxford University Press.

Do, K.-A. (2002). Biostatistical approaches for modeling longitudinal and event time data, *Clin. Cancer Res.* **8**(8): 2473–2474.

Doksum, K. A. and Gasko, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis, *International Statistical Review* **58**: 243–252.

Duchesne, T. (1999). *Multiple Time Scales in Survival Analysis*, PhD thesis, University of Waterloo.

Duchesne, T. and Lawless, J. (2000). Alternative time scales and failure time models, *Lifetime Data Analysis* **6**(2): 157–179.

Efron, B. (2002). The two-way proportional hazards model, *Journal of the Royal Statistical Society, Series B, Methodological* **64**(4): 899–909.

Farewell, V. T. and Cox, D. R. (1979). A note on multiple time scales in life testing, *Applied Statistics* **28**: 73–75.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach, *Statistics in Medicine* **15**: 1663–1685.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*, Wiley.

Fleming, T. and Lin, D. (2000). Survival analysis in clinical trials: Past developments and future directions, *Biometrics.* **56**(4): 971–983.

Gentleman, R. and Vandal, A. C. (2002). Nonparametric estimation of the bivariate CDF for arbitrarily censored data, *The Canadian Journal of Statistics* **30**(4): 557–571.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares, *Biometrika* **73**: 43–56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation, *Biometrika* **76**: 622–623.

Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**: 515–526.

Guo, G. and Rodríguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala, *Journal of the American Statistical Association* **87**: 969–976.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages, *The American Statistician* **58**: 16–24.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapmn and Hall, London.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models, *Stat. Sci.* **1**: 297–318.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics* **1**: 465–480.

Holford, T. (1998). Age-period-cohort analysis, *in* P. Armitage and T. Colton (eds), *Encyclopedia of Biostatistics*, John Wiley and Sons, Ltd.

Hosmer, D. and Lemeshow, S. (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*, John Wiley and Sons, Inc.

Hougaard, P. (1995). Frailty models for survival data, *Lifetime Data Analysis* **1**: 255–273.

Huber, P., Ronchetti, E. and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models, *J. R. Statist. Soc. B* **66**: 893–908.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*, Springer-Verlag Inc.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**(3): 299–314.

Ivanoff, B. and Merzbach, E. (2002). Random censoring in set-indexed survival analysis, *The Annals of Applied Probability* **12**: 944–971.

Jewell, N. and Kalbfleisch, J. (1996). Marker processes in survival analysis, *Lifetime Data Analysis* **2**: 15–29.

Johansen, S. (1983). An extension of Cox's regression model, *International Statistical Review* **51**: 165–174.

Jones, M. P. and Crowley, J. (1992). Nonparametric tests of the Markov model for survival data, *Biometrika* **79**: 513–522.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, second edn, John Wiley and Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**: 457–481.

Keiding, N. (1990). Statistical inference in the lexis diagram, *Phil. Trans. R. Soc. London A* **332**: 487–509.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology, *Statistics in Medicine* **18**: 2353–2363.

Kevane, M. and Levine, D. I. (2003). Changing status of daughters in indonesia, *Paper c03-126*, Center for International and Development Economics Research. University of California, Barkeley. http://Repositories.Cdlib.Org/Iber/Cider/C03-126.

Korn, E., Graubard, B. and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale, *Am-J-Epidemiol* **145**: 72–80.

Kuczmarski, R., Ogden, C. and Guo, S. (2002). CDC growth charts for the united states: Methods and development., *Vital Health Stat 11 246*, National Center for Health Statistics.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**: 963–974.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models: A synthesis of generalised linea models, random-effet models and structure dispersions, *Biometrika* **88**: 987–1006.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrika.* **73**: 13–22.

Liestøl, K. and Andersen, P. (2002). Updating of covariates and choice of time origin in survival analysis: Problems with vaguely defined disease states, *Statist. Med.* **21**: 3701–3714.

Lin, H., McCulloch, C. E. and Mayne, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables, *Statistics in Medicine* **21**(16): 2369–2382.

Lin, H., Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer, *Journal of the American Statistical Association* **97**(457): 53–65.

Lind, T. (2004). *Iron and Zinc in Infancy: Results from Experimental Trials in Sweden and Indonesia*, Umeå university medical dissertations, Epidemiology and Public Health Sciences, Department of Public Health and Clinical Medicine, and Pediatrics Department of Clinical Sciences, Umeå University, Sweden.

Lindkvist, M. (2000). *Added Variable Plots and Influence in Cox's Regression Model.*, PhD thesis, Department of Statistics, Umeå University.

Machfudz, S. (1998). *Effect of Morbidity on Change in Mid-upper-arm Circumference in Children Under Five Years of Age. a Cohort Study in Purworejo, Central Java, Indonesia*, Master's thesis, Department of Epidemiology and Public Health Umeå University.

Manda, S. (2001). A comparison of methods for analysing a nested frailty model to child survival in malawi, *Australian New Zealand Journal of Statistics* **43**(1): 7–16.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second Edition)*, Chapman & Hall Ltd.

Mosley, W. and Chen, L. (1984). An analytical framework for the study of child survival in developing countries, *Population and Development Review* **10**: 25–48. Suppl.

Ng, E. T. M. and Cook, R. J. (1997). Modeling two-state disease processes with random effects, *Lifetime Data Analysis* **3**: 315–335.

Oakes, D. (1995). Multiple time scales in survival analysis, *Lifetime Data Analysis* **1**: 7–18.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS, *Journal of the American Statistical Association* **88**: 719–726.

Pearce, N. (1992). Methodological problems of time-related variables in occupational cohort studies, *Rev Epidemiol Sante Publique* **40 Suppl 1**: S43–54.

Pebley, A. and Stupp, P. (1987). Reproductive patterns and child mortality in Guatemala, *Demography* **24**(1): 43–60.

Prentice, R. (1982). Covariate measurement errors and parameter estimates in a failure time regression model., *Biometrika* **69**: 331–342.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria, *Statistics in Medicine* **8**: 431–440.

Prentice, R. L. (1999). On non-parametric maximum likelihood estimation of the bivariate survivor function, *Statistics in Medicine* **18**: 2517–2527.

R Development Core Team (2004). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. 3-900051-00-3.
*`http://www.R-project.org`

Rabe-Hesketh, S., Yang, S. and Pickles, A. (2001). Multilevel models for censored and latent responses, *Stat. Methods Med. Res.* **10**: 409–427.

Rice, A., Sacco, L., Hyder, A. and Black, R. (2000). Malnutrition as an underlying cause of childhood deaths associated with infectious diseases in developing countries, *Bulletin of the World Health Organization* **78**: 1207–1221.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect, *Mathematical Modelling* **7**: 1393–1512.

Rochon, J. and Gillespie, B. (2001). A methodology for analysing a repeated measures and survival outcome simultaneously., *Stat.Med.* **20**(8): 1173–1184.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil, *Journal of the American Statistical Association* **92**: 426–435.

Scrimshaw, N. S. (2003). Historical concepts of interactions, synergism and antagonism between nutrition and infection, *J. Nutr.* **133**: 316S–321S.

The Cebu Study Team (1991). Underlying and proximate determinants of child health: The cebu longitudinal health and nutrition study, *Am. J. Epidemiol* **133**: 185–201.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag Inc.

Trussell, J. and Hammerslough, C. (1983). A hazard-model analysis of the covariates of infant and child mortality in sri lanka, *Demography* **20**: 1–26.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview, *Statistica Sinica* **14**: 809–834.

Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS, *Journal of the American Statistical Association* **90**: 27–37.

UNICEF (2003). Child Survival and Health. `http://www.childinfo.org/eddb/health.htm`. Accessed October 13, 2003.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*, Springer-Verlag, Inc.

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**: 439–454.

Wahab, A., Winkvist, A., Stenlund, H. and Wilopo, S. (2001). Infant mortality among Indonesian boys and girls: Effect of sibling status, *Annals of Tropical Paediatrics* **21**(1): 66–71.

Wibowo, T. (2000). *Does Poor Nutritional Status Lead to Morbidity? A Longitudinal Study of Infants 6 - 12 Months in Purworejo, Central java, Indonesia*, Master's thesis, Department of Epidemiology and Public Health Umeå University.

Wilopo, S. and CHN-RL Team (1997). *Key Issues on Research Design, Data Collection and Management. Community Health and Nutrition Research Laboratory, Faculty of Medicine, Gadjah Mada University*, Reprint Series No. 2, Community Health and Nutrition Research Laboratory, Yogyakarta.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics* **53**: 330–339.

Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events, *Applied Statistics* **50**(3): 375–387.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach, *Journal of the American Statistical Association* **86**: 79–86.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**: 121–130.

Zeger, S. L. and Liang, K.-Y. (1991). Feedback models for discrete and continuous time series, *Statistica Sinica* **1**: 51–64.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics* **44**: 1049–1060. (Correction: V45 P347).

Zohoori, N. and Savitz, D. (1997). Econometric approaches to epidemiologic data: Relating endogeneity and unobserved heterogeneity to confounding, *Ann. Epidemiol* **7**: 251–257.

Appendix

A-1 Simulating alternative time scale

The simulation procedure for the alternative time scales in Section 4.4.1 is described here. The true duration T is generated by the ordinary Cox model

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta Z),\ t > 0, \tag{A-1}$$

where $\lambda(t \mid Z)$ is the hazard for an individual, $\lambda_0(t)$ is the baseline hazard, parametrically specified in this simulation, Z is a zero-one fixed time covariate with coefficient β.

Z is specified by the Bernoulli distribution with probability 0.4 of success and the true value of β is 2. The baseline hazards are specified by Gompertz, exponential and Weibull hazard functions. Table A-1 shows the detailed specifications.

Table A-1: The specification of hazard functions and times T generation

Baseline	hazard	T generation	specification
Gompertz	$\lambda_0(t) = \theta_1 e^{\theta_2 t}$	$T = \frac{1}{\theta_2} \log(-\frac{\theta_2}{\theta_1} \frac{\log(u)}{\Psi_i} + 1)$	$\theta_1 = 0.15$, $\theta_2 = 2$
exponential	$\lambda_0(t) = \theta$	$T = -\frac{\log(u)}{\theta \Psi_i}$	$\theta = 0.85$
Weibull	$\lambda_0(t) = \theta_1 \theta_2 (\theta_2 t)^{\theta_1 - 1}$	$T = \frac{1}{\theta_2} (\frac{-\log(u)}{\Psi_i})^{1/\theta_1}$	$\theta_1 = 1.2$, $\theta_2 = 0.5$

$\Psi_i = \exp(\beta Z_i)$, $\quad u \sim U(0, 1)$

After T is generated, T_1 and T_2 are generated by adding δ_1 and δ_2, respectively. In the simulation, δ_1 is $U(0, 1)$ or exponential(0.5); δ_2 is $U(0.5, 2)$ or exponential(1.25). Samples of size $n = 200$ individuals were generated according to this procedure with 1000 replications.

A-2 Simulating dual time scales

The simulation procedure for the dual time scales in Section 4.4.2 used time-dependent covariate models. In general, if we have a Cox model with time dependent covariate

$$\lambda(t \mid Z(t)) = \lambda_0(t)\Psi(\beta, t),\ t > 0, \tag{A-2}$$

the duration T can be generated through the relationship between hazard and survival. If T has distribution function $F(t)$ or survival function $S(t)$ then $U = F(T)$ or similarly $U = S(T)$ will follow a uniform $U(0,1)$ distribution.

Under model (A-2) the cumulative hazard function for T is

$$\begin{aligned} G(t) &= \Lambda(t \mid Z(s), 0 \leq s \leq t) \\ &= \int_0^t \lambda_0(y)\Psi(\beta, y)dy \end{aligned} \tag{A-3}$$

so that

$$S(t) = \exp(-G(t)). \tag{A-4}$$

Now, $U = S(T)$ is $U(0,1)$. Therefore, solving $U = \exp(-G(T))$ for T gives what we want.

Suppose T has hazard function

$$\lambda(t \mid Z(t+\delta)) = \lambda_0(t)\Psi(\beta, t+\delta),\ t > 0, \tag{A-5}$$

where $\lambda(t \mid Z(t+\delta))$ is the hazard function for an individual the covariate process Z, $\lambda_0(t)$ is the baseline hazard, parametrically specified in this simulation, and $\Psi(\beta, t)$ is specified as

$$\Psi(\beta, t) = \exp(\beta_1\eta + \beta_2(t+\delta)),\ t > 0, \tag{A-6}$$

where β_1 and β_2 are parameters specified in the simulation, and η and δ follow certain distributions.

The dual times T_1 and T_2 can be generated from model (A-5) after specifying the baseline hazard function λ_0. In this simulation, we specify a constant hazard θ such that (A-4) has a closed form solution,

$$\lambda(t \mid Z(t)) = \theta\exp(\beta_1\eta + \beta_2(t+\delta)),\ t > 0. \tag{A-7}$$

The cumulative hazard function for an individual with covariate process Z is

$$
\begin{aligned}
G(t) &= \Lambda(t \mid Z(s), 0 \leq s \leq t) \\
&= \int_0^t \theta \exp(\beta_1 \eta + \beta_2(y+\delta))dy \\
&= \theta e^{\beta_1\eta+\beta_2\delta}\left[\frac{e^{\beta_2 y}}{\beta_2}\right]_{y=0}^{t} \\
&= \theta e^{\beta_1\eta+\beta_2\delta}\left[\frac{e^{\beta_2 t}-1}{\beta_2}\right]. \qquad \text{(A-8)}
\end{aligned}
$$

In the simulation study, we specify a constant hazard $\theta = 1.2$, the true coefficients $\beta_1 = 1.5$, $\beta_2 = 0, 1$, zero-one fixed covariate $\eta \sim \text{Bernoulli}(p = 0.45)$, and δ follows exponential with rate 0.85 and $U(0,2)$. Using this specification T_1 and T_2 can be generated through the inverse of G,

$$
G^{-1}(y) = \begin{cases} \frac{1}{\beta_2}\log\left(\frac{\beta_2 y}{\theta e^{\beta_1\eta+\beta_2\delta}}+1\right) & \text{for } \beta_2 \neq 0 \\ \frac{y}{\theta}e^{-\beta_1\eta} & \text{for } \beta_2 = 0 \end{cases} \qquad \text{(A-9)}
$$

and $T_1 = G^{-1}(-\log(u))$ with $u \sim U(0,1)$; $T_2 = T_1 + \delta$. Samples of size $n = 200$ individuals were generated according to this procedure with 1000 replications.

A-3 Simulating longitudinal measurements and event-time data

The simulation method in Section 5.4 uses the same principle as in A-2, in which the event times are generated through the inverse of the cumulative hazard function. However, in this simulation a longitudinal model is involved.

A-3.1 Time-dependent covariate model

This simulation is based on Equations (5.1), (5.2), and (5.3) (Section 5.2).

Specifically we have the longitudinal growth curve model

$$Y_i^\star(t) = (\alpha_1 + a_{1i}) + (\alpha_2 + a_{2i})t, \quad t > 0,\ i = 1, \ldots, n, \tag{A-10}$$

where $Y_i^\star(t)$ are longitudinal measurements. The random coefficients a_{1i} and a_{2i} are assumed to follow a bivariate Gaussian distribution with mean zero and variance-covariance matrix $\mathbf{\Sigma}$.

The measurements are made intermittently for each individual i and with error, therefore the simulated model for the growth curve is

$$Y_{ij} = Y_i^\star(t_{ij}) + \epsilon_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{A-11}$$

where t_{ij}, $i = 1, \ldots, n$, $i = 1, \ldots, m$ are time points of measurement. The measurement errors ϵ_{ij} are assumed to be mutually independent Gaussian distributed with mean zero and variance σ_ϵ.

The hazard function is modeled as a Cox model with constant baseline hazard

$$\lambda_i(t) = \theta \exp(\beta_1 Z_i + \beta_2 Y_i^\star(t)),\ t > 0,\ i = 1, \ldots, n. \tag{A-12}$$

Substituting $Y_i^\star(t)$ from Equation (A-10) and dropping the index i, the cumulative hazard of (A-12) can be written as

$$G(t) = K \frac{\exp\left(\beta_2(\alpha_2 + a_2)t\right) - 1}{\beta_2(\alpha_2 + a_2)},\ t > 0, \tag{A-13}$$

where $K = \theta \exp(\beta_1 Z + \beta_2 \alpha_1 + \beta_2 a_1)$.

The event times are generated by $G^{-1}(-\log(u))$ with $u \sim U(0, 1)$ (see (A-9)). Since the simulation is for repeated events, for one individual we assume that the inter event times are generated by the same model but the time origin is advanced by a certain random amount after each event time. In the context of morbidity, we call the advancing of the time origin as duration of illness. For this simulation we choose the lognormal distribution as the distribution of illness duration.

In the simulation, we specified the parameters for the hazard model as $\theta = 0.4$, $\beta_1 = 1.2$ and varied $\beta_2 = 0, -0.1$, illness duration was lognormal(0, 0.3); and in the growth curve model, we used the parameter values $\alpha_1 = 6.5$,

$\alpha_2 = 0.17$, $\sigma_\epsilon = 0.2$, and

$$\mathbf{\Sigma} = \begin{pmatrix} 0.9 & -0.04 \\ -0.04 & 0.01 \end{pmatrix}.$$

These specified values are roughly equal to the parameter estimates obtained from the ZINAK study especially for the weight growth model. Age time scale is used in the simulation starting from 6 to 12 months, which is also roughly the same as in the ZINAK study. The counting process style input `(start`, `stop]`, `event` is used for the repeated events.

The longitudinal measurements were generated at some defined time intervals. The measurements time points were t_{i1}, t_{i2}, t_{i3} and were not exactly the same for all individuals. This was done by adding a random uniform $U(-0.4, 0.4)$ to time points 6, 9, 12 for each individual. Samples of size $n = 50$ individuals were generated according to this procedure with 500 replications.

A-3.2 Joint model

Simulation of the joint model is based on Equations (5.1), (5.2) and (5.4) (Section 5.2). The procedure for the simulated longitudinal measurements is similar to that of the time-dependent covariate model with the following modification

$$Y_i(t) = (\alpha_1 + a_{1i}) + (\alpha_2 + a_{2i})t + \alpha_3 Z_i + \epsilon_i, \ t > 0, \ i = 1, \ldots, n, \quad \text{(A-14)}$$

where now we have Z_i in the model.

The simulated event-times were generated from the hazard function

$$\lambda_i(t) = \theta \exp(\beta_1 Z_i + \beta_2(a_{1i} + a_{2i}t)), \ t > 0. \quad \text{(A-15)}$$

The cumulative hazard of (A-15) is

$$G(t) = K \frac{\exp\left(\beta_2 a_2 t\right) - 1}{\beta_2 a_2}, \ t > 0, \quad \text{(A-16)}$$

where $K = \theta \exp(\beta_1 Z + \beta_2 a_1)$. The event times are then generated by $G^{-1}(-\log(u))$ with $u \sim U(0, 1)$.

The duration of illness, θ, β_1, σ_ϵ and $\mathbf{\Sigma}$, as well as the schedule of measurement times t_{ij} were specified similarly as in the time-dependent covariate model. The α's were specified as $\alpha_1 = 6.5$, $\alpha_2 = 0.5$, $\alpha_3 = 1.5$ and varied $\beta_2 = 0, 1$. Samples of size $n = 50$ individuals were generated according to this procedure with 500 replications.

Statistical Studies
issued by
Department of Statistics, Umeå University
SE–901 87 Umeå, Sweden

1. Gustafsson, Lennart: Några aspekter på stickprovsteorier vid ändliga populationer med tillämpningar på tvåstegsurval (1968).

2. Pollak, Kay: Variationsskattningar baserade på kvadratiska former av ordnade variabler, några illustrationer (1969).

3. Cassel, Claes-Magnus: Inferensproblemet vid ändliga populationer, några synpunkter (1970).

4. Wretman, Jan-Håkan: Om inferens vid ändliga populationer under superpopulationsantagande (1970).

5. Carlsson, Olle: Om fördelningen av en summa av vägda oberoende Poissonvariabler med tillämpningar inom statistisk inferensteori och stokastiska processer (1970).

6. Stenlund, Hans och Westlund, Anders: A Monte-Carlo Study of Some Sampling Designs (1974).

7. Westlund, Anders: Estimation and Prediction Interdependent Systems in the Presence of Specification Errors (1975).

8. Björnham, Åke och Wiklund, Dan-Erik: Analysis of Fetal Heart Rate Variability During Labour: Registration, Estimation, and Decision (1976).

9. Hållberg, Bengt: Statistiska modeller för banbrottsfrekvens hos tryckpapper (1976).

10. Freij, Lennart och Wall, Stig: Exploring Child Health and its Ecology (1977).

11. Baudin, Anders: On the Application of Short-term Causal Models (1977).

12. Brännäs, Kurt: On Estimation in Economic System in the Presence of Time Varying Parameters (1980).

13. Nyquist, Hans: Recent Studies on Lp-Norm Estimation (1980).

14. Törnkvist, Birgitta: Quantifying Structural Change - A Model Based Approach (1988).

15. Laitila, Thomas: Estimation in Truncated and Censored Regressions (1989).

16. Carlsson, Olle: On Quality Selection (1990).

17. Segerstedt, Bo: On Conditioning and Ridge Estimation in Generalized Linear Models (1991).

18. Öhman, Marie-Louise: Contributions to Generalized Wilcoxon Rank Tests (1992).

19. Wiklund, Stig-Johan: Control Charts and Process Adjustments (1994).

20. Arnoldsson, Göran: Generalised Linear Models and Optimal Design (1994).

21. Öhman, Marie-Louise: Aspects of Analysis of Small-Sample Right Censored Data Using Generalized Wilcoxon Rank Tests (1994).

22. Arnoldsson, Göran: Optimal Design for Inference in Generalized Linear Models (1997).

23. Bränberg, Kenny: On Test Score Equating (1997).

24. Häggström, Jonas: The Minimax Approach to Optimum Design of Experiments (2000).

25. Lindkvist, Marie: Added Variable Plots and Influence in Cox's Regression Model (2000).

26. Pettersson, Hans: Optimum in Average and Minimax Designs for Estimation of Generalized Linear Models (2001).

27. Häggström Lundevaller, Erling: Tests of Random Effects in Linear and Non-Linear Models (2002).

28. Adler, John: Statistical Models for Estimating Career Mobility (2003).

29. Wiberg, Marie: Computerized Achievement Tests - Sequential and Fixed Length Tests (2003).

30. Danardono: Event History Analysis of Childhood Mortality and Morbidity in Purworejo, Indonesia (2003).

31. Puu, Margareta: Optimum Experimental Designs for Generalized Linear Models with Multinomial Response (2003).

32. Appelgren, Jari: Locally D-optimal Designs for Bivariate Logistic Regression (2004).

33. Danardono: Multiple Time Scales and Longitudinal Measurements in Event History Analysis (2005).