Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Scuola di Dottorato in
Scienze Statistiche
Ciclo XX

Theoretical and applied issues arising from the joint modelling of longitudinal response processes and time to competing events

Daniela Zugna

Direttore: Prof.ssa A. Salvan

Supervisore: Prof. N. Torelli

Co-supervisori: Prof.ssa B. De Stavola, Prof.ssa C. Di Serio

31 gennaio 2008

Riassunto

Negli studi osservazionali che hanno come obiettivo l'analisi dell'HIV e dell'AIDS, i pazienti sono seguiti attivamente dalla loro entrata nello studio fino al verificarsi di un evento (quali l'AIDS, la morte, o la conclusione dello studio stesso) e sono soggetti a frequenti visite nelle quali vengono rilevati dati clinici, biologici e riguardanti la tipologia di trattamento ai quali sono sottoposti. I dati raccolti si riferiscono alla misurazione dei biomarcatori della progressione della malattia, delle terapie assunte e dello stato di salute generale dei pazienti. L'analisi statistica dei dati provenienti da tali studi si scontra con la necessità di affrontare molteplici problemi causati dalla natura longitudinale e di sopravvivenza dei dati. Infatti, a causa della natura osservazionale dei dati, le visite cliniche avvengono ad intervalli irregolari di tempo, in tempi diversi ed in numero diverso per ogni partecipante allo studio. Inoltre, le misurazioni dei biomarcatori possono essere soggette ad errori di misurazione e/o assumere valori anomali determinati da particolari condizioni fisiologiche contingenti. Infine, a causa della complessità del processo sottostante la progressione della malattia, particolare attenzione deve essere rivolta ai tempi in cui si verificano gli eventi di carattere clinico, quali l'insorgenza di malattie associate all'AIDS e la morte, la cui occorrenza dipende non solo dal tempo trascorso dalla seroconversione, ma anche da ulteriori fattori quali l'età e la somministrazione di nuove terapie. Un altro importante aspetto da non trascurare riguarda la censura informativa. Infatti, in aggiunta alla censura non informativa determinata dalla conclusione dello studio, forme di censura dipendente potrebbero essere dovute allo stato di salute del paziente, situazione che si verifica, ad esempio, quando i pazienti più malati sono portati ad abbandonare lo studio.
L'obiettivo di questo lavoro consiste nell'analisi di dati epidemiologici relativi ai pazienti che hanno contratto l'HIV. I dati oggetto di studio provengono da un progetto organizzato dalla CASCADE (Concerted Action on SeroConversion to

AIDS and Death in Europe) e rappresentano uno dei più grandi studi multicentrici sull'AIDS, risultato di una collaborazione che rappresenta 22 coorti in Europa, Australia e Canada. A differenza di altri studi in questo campo, la data di sieroconversione è nota per tutti i partecipanti. Inoltre, i due principali biomarcatori della progressione della malattia, i CD4 e la carica virale, sono registrati longitudinalmente dall'entrata fino all' uscita del paziente dallo studio e con essi tutti gli eventi associati all'AIDS. Grazie alla completezza ed all' affidabilità dei dati a disposizione, è stato possibile proporre modelli alternativi per la modellizzazione congiunta dei dati longitudinali, il CD4 e la carica virale, e degli eventi caratterizzanti la storia del paziente. Per una maggiore semplicità, l'analisi è ristretta ad un gruppo di 1090 uomini che hanno contratto l'HIV tramite rapporti omosessuali.
La domanda d'interesse alla quale si è voluto rispondere, trae origine da un dibattito recente sul tempo ottimale in cui somministrare la terapia HAART (highly active antiretroviral therapy) ad individui infetti da HIV. Dall'introduzione dell'HAART nel 1995, lo scenario della ricerca sull'HIV è stato profondamente alterato. Questo trattamento ha, infatti, portato ad una sostanziale riduzione della mortalità e della progressione della malattia, sopprimendo la carica virale ed aumentando i CD4. Il dibattito nasce dal fatto che, mentre è risaputo che la terapia deve essere somministrata ai pazienti in cui i CD4 raggiungono un valore minore di 200 cellule/mm^3, ci sono ancora opinioni discordanti sui vantaggi dovuti al deferire la somministrazione della terapia fino a quando i CD4 raggiungono un valore minore o uguale a 350 cellule/mm^3 piuttosto che le originalmente raccomandate 500.
Questa tesi propone un metodo atto a modellare congiuntamente i dati longitudinali ed i rischi competitivi. Modellando il processo longitudinale dei due biomarcatori attraverso un modello bivariato ad effetti causali, si è analizzata la loro dipendenza attraverso gli effetti casuali e sono stati risolti i problemi causati dalle irregolarità e dagli errori di misurazione. Inoltre, considerando la censura informativa come un rischio competitivo dipendente ed utilizzando un modello a rischi proporzionali, si potuto ottenere stime non distorte del processo longitudinale. Allo stesso tempo, attraverso i parametri che specificano l'associazione tra il processo longitudinale e quello di sopravvivenza, è stato possibile modellare e valutare l'effetto dei biomarcatori, corretto per ulteriori variabili dette di "confondimento", sui rischi competitivi. La caratteristica principale della modellizzazione congiunta è che i parametri che descrivono il processo longitudinale e quelli che descrivono quello di sopravvivenza, espresso in funzione del processo longitudinale, sono stati stimati simultaneamente, utilizzando così in maniera più efficiente i dati.

Abstract

The statistical analysis of observational data arising from HIV/AIDS research is generally faced with complexities that arise from both the longitudinal and survival features of the data. In this field patients are actively followed up from entry into the study till either an AIDS-related illness or death, with frequent follow-up visits where several clinical, biological and treatment data are collected. Thus the available information includes records on biomarkers of the progression of disease, changing treatments, and disease/survival status. Because of the observational nature of the data, the follow-up visits occur at irregular time intervals, usually at varying time points and in unequal numbers for different study participants. In addition, the "true" level of each of these biomarkers may be measured with error because of laboratory and/or physiological variations. Further, because of the complex process underlying this disease, time to several event types, such as AIDS-related events, besides time to death, are of interest, with their occurrence depending not only on the time elapsed since initial seroconversion, but also on concomitant or intervening factors such as aging and initiation of various treatments. One other important issue that affects these data, possibly more than any other type of survival data, is informative censoring: beside the naturally occurring censoring of the follow-up due to the study closure, other sources of censoring occur via withdrawal from the study, due for example to poor health. In summary, in this field data have both a longitudinal and a survival component. The longitudinal data consist of irregularly and possibly poorly measures of important biomarkers, while the survival data involve multiple end-points, multiple time-scale and are likely to be affected by informative censoring which, if ignored, leads to biased results.

We have the opportunity to analyse epidemiological data on HIV patients arising from one of the largest AIDS multicentre studies, the CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe) Study, a collaboration

representing 22 cohorts based in Europe, Australia and Canada. Unlike other studies in this field, the date of seroconversion of all participants is reliably estimated. In addition both their CD4 cell counts and RNA viral load are recorded longitudinally from entry into the study till the end of their follow-up (due to censoring or death), and all AIDS-related events are carefully recorded. We can therefore take advantage of the quality of these data to explore alternative models for the joint distribution of the longitudinal data on CD4 cell count and RNA viral load and on the event/survival data. For simplicity we will focus our analyses on the subgroup of all 1090 homosexual men who are part of CASCADE. Our aim is motivated by the current debate regarding the optimal timing of initiation of highly active antiretroviral therapy (HAART) in chronically HIV-infected individuals. This is an important clinical question because the landscape of HIV research has been profoundly altered since the introduction of HAART in 1995. This treatment led to a substantial reduction in mortality and disease progression to AIDS by suppressing HIV viral load and increasing CD4 cell counts. There is a shift in opinion toward deferring HAART initiation until CD4 cell counts fall below 350 cell/μl, rather than the originally recommended 500 cell/μl, while there is general agreement that HAART should definitely be prescribed to patients whose CD4 cell counts are less than 200μl.

This thesis proposes a methodology for modelling the joint variation over time of the two biomarkers and of the survival processes of a set of competing events. Modelling two longitudinal response processes as a bivariate linear mixed effects model, with knots at relevant times, will account for the dependence between two biomarkers by random effects while overcoming the problem of irregularly measured data and of the possible measurement errors. Furthermore modelling the informative withdrawals from the study as dependent competing risks, by estimating the so called cumulative incidence functions whithin a proportional hazards model, will allow for an unbiased estimate of the markers' processes. At the same time the parameters that specify the association between the markers processes and the survival processes will allow to model the effect of the biomarkers, adjusted for other covariates, on the competing events. The essential feature of joint modelling is that the parameters which describe the longitudinal response processes and those which describe the failure risks, as a function of the longitudinal processes, are estimated simultaneously, making a more efficient use of data.

Contents

Part I

Aims and outlines

1

Introduction

1.1 Motivating example

In many epidemiologic studies, a cohort of subjects is followed over time to investigate the relationship between one or more explanatory variables and the risk of developing a disease. When it is possible to identify a disease marker, firstly it is of interest to monitor its progression and its dependence on the other variables, then to understand how the marker's pattern is related to the disease risk (Jewell and Kalbfleisch, 1996). Such setting gives raise to several challenges, which derive from both the longitudinal and the survival features of data.
Observational studies in HIV/AIDS[1] research constitute a typical example of this type of setting: immunologic and virologic markers are measured repeatedly over time on each patient, and the patients are actively followed up from entry into the study till either development of an opportunistic infection associated with HIV disease or death, depending on the survival endpoint of interest. In this case, immunologic and virologic markers might be used as time-varying predictor variables, or as markers.
The most common indicator used to evaluate the immunological status of an HIV patient is the CD4+lymphocyte count, which measures the number of CD4 cells in each cubic millimetre of blood. A normal count in a healthy, HIV-negative adult can vary but is usually between 500 and 1500 cells/mm^3. CD4 cells are a type of lymphocyte that co-ordinate the immune system's response to certain micro-organisms such as viruses, the higher the CD4+ cell count is, the lower the risk of infection is. Viral load is a measure of the severity of a viral infection, and is measured by estimating the amount of virus in the blood plasma, for example, reported as the number of RNA "copies" in a millilitre of blood. The test can detect 50 copies at

[1] HIV: human immunodeficiency virus; AIDS: acquired immunodeficiency syndrome

the lower end, 10_4 at the upper end. A high viral load, i.e. greater than 100000 copies, indicates a higher risk of disease progression of HIV patient, while a low viral load, i.e. less than 10000 copies, indicates that in the near future risk of disease progression is relatively low.

The field of HIV research has been profoundly altered since the introduction of HAART[2] in 1995. HAART is defined as three or more antiretroviral drugs containing at least two nucleoside reverse transcriptase inhibitors (NRTIs) plus a protease inhibitor (PI), a nonnucleoside reverse transcriptase inhibitor (NNRTI), or abacavir (ABC). This treatment has led to a substantial reduction in mortality and disease progression to AIDS by suppressing viral load and increasing CD4 cell count. Yet, a debate regarding the optimal timing of initiation of HAART in chronically HIV-infected individuals is in progress (Phillips *et al.*, 2001; Pomerantz, 2001, Lepri *et al.*, 2001; Grant *et al.*, 2003). There is a shift in opinion toward deferring HAART initiation until CD4 cell count falls below 350 cell/μl (Carpenter *et al.*, 2000), rather than the originally recommended 500 cell/μl (Pomerantz, 1995; Opravil *et al.*, 2002), while there is general agreement that HAART should definitely be prescribed to patients whose CD4 cell count is lower than 200μl. The main arguments for starting HAART early is to reduce the risk of opportunistic infections, with the consequent improvement in quality of life, and to preserve HIV-specific cellular response, which seems to be better if the therapy is started when the CD4 cell count is high (Ledergerber *et al.*, 1999a; Ledergerber *et al.*, 1999b). In contrast, while the short- and medium-term side-effects of HAART are known, the long-term toxicity is not, and therefore the benefits of delayed treatment would include instead avoiding the side-effects of lifelong treatment with antiretroviral drugs and minimizing the development of viral drug resistance (Molla *et al.*, 1996; Miller *et al.*, 1999).

The clinical question our method wants to dress is inspired by this debate. We analyse data on HIV patients arising from one of the largest AIDS multicentre studies, the CASCADE[3] Study, and compare the disease progression, expressed by biomarkers' pattern and the event histories, of three groups of individuals defined according to CD4 cell count at HAART initiation, lower than 200 cell/μl, included between 200 and 350 cell/μl, and higher than 350 cell/μl, respectively. Specifically, we are interested in investigating the different elapsed times between seroconversion and HAART initiation, the biomarkers' pattern, and the risk of failure from AIDS, the interruption of therapy or of changing therapy in terms of the biomarkers pattern and of other covariates, for each group.

[2] HAART: highly active antiretroviral therapy

[3] Concerted Action on SeroConversion to AIDS and Death in Europe

1.2 Statistical issues

Randomized clinical trials could be designed to determine the optimal stage of HIV infection to initiate therapy. One such trial would randomize patients whose CD4 cells count is between 350-500/μl to either immediate or deferred treatment arms and monitor the disease progression. The principal strength of such a trial would be to minimize the effect of subjective factors related to treatment decisions. Yet the comparability could be compromised because of the lengthy follow-up, during which therapies are likely to change, and overall the differences in AIDS risk might be due to treatment choices and not to the timing of treatment. On the other hand observational studies need to account for the lack of randomization in treatment assignment and for the lead time, which is the additional survival time required by those who start therapy at later stage in order to progress from the early stage to time of initiation therapy. For example, a person initiating HAART with a CD4 cell count of 200 cells/μl is likely to develop AIDS more quickly than a person initiating HAART with a value of 350μl cells. However the time it took the first individual to reach a level of 200 starting from 350μl cells has to be considered. Different approaches to account for lead time have been proposed (Grant *et al.*, 2003; Cole *et al.*, 2004). The basic idea consists of selecting a common time-scale for the patients classified according to their CD4 counts at initiation of therapy. Since in CASCADE the date of seroconversion of all participants is reliably estimated, unlike other studies in this field, we can deal with the "lead time bias", by choosing the time since seroconversion as the time-scale for survival analysis and by taking to account the elapsed time between seroconversion and HAART initiation when modelling the biomarkers' pattern and the failure risks. Performing the analysis on time since seroconversion also allows to control the potential confounding effect of CD4 levels at start of treatment on the effect of the treatment itself. In figure 1.1 the different time-scales, which could be used to analyse these data, time since seroconversion and time since HAART initiation are compared. It is visible the loss of information in the second case, as the disease duration before treatment initiation is ignored.

In CASCADE both CD4 cell count and viral load are recorded longitudinally from entry into the study till the end of the patients' follow-up. CD4 cell count are measured by flow cytometry. For viral load quantification, the Amplicor HIV-1 Monitor Tests or Quantiplex HIV-1 RNA assays are the most often used. Before HIV RNA assays were developed in mid-1990s, CD4 cell count served as primary biomarker of progression of HIV. Later the combination of these two markers was shown to be more predictive to clinical outcomes. In general, it is believed that the virologic response, measured by viral load, and immunological response, measured by CD4

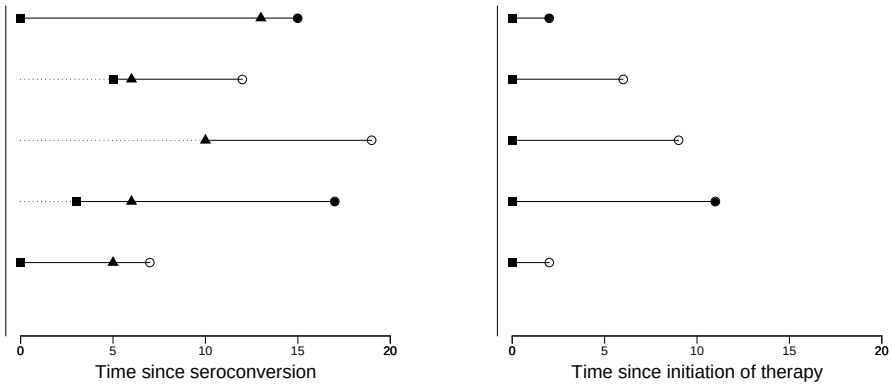

Fig. 1.1. Plots of follow-up of five individuals, depending on time-scale: time since seroconversion and time since HAART initiation, respectively. Time of seroconversion (starting value), time of entry into the study (square), time of HAART initiation (triangle), time of loss to follow-up (empty point), time of exit from the study (full point).

cell counts, are negatively correlated during the submission to therapy and their relationship is not costant over the time. However the response of the two biomarkers to therapy is not clear yet, i.e. when a patient starts a successfull treatment regimen, the viral load tends to drop drastically, while the CD4 cell count may take longer to respond or may not respond at all. The decrease in HIV-RNA viral load and the corresponding increase of CD4 cells count, because of recovering of the immune system, are not always observable. Hence there is considerable interest in monitoring these biomarkers' development and their correlation over time since seroconversion, and in understanding how these markers, together with other covariates, may influence the incidence rates of all the events of clinical interest.
Because of the observational nature of the data, the biomarkers are measured at irregular time intervals, usually at varying time points and in unequal numbers for different study participants, as shown in figure 1.2. These markers may be prone to measurement error and high within patient variability due to biological fluctuations (Hoover *et al.*, 1992). Because of difficulty and high cost of assays to quantify the viral load, the viral load measures are less numerous and more variable than CD4 cell counts. Furthermore, viral load measures could be left censored, mainly after therapy initiation, when the biomarker undergoes a drastic fall, since most of the assays that exist to quantify the viral load in blood are characterized by a low quantification limit. Modelling the biomarkers' pattern provides estimates for

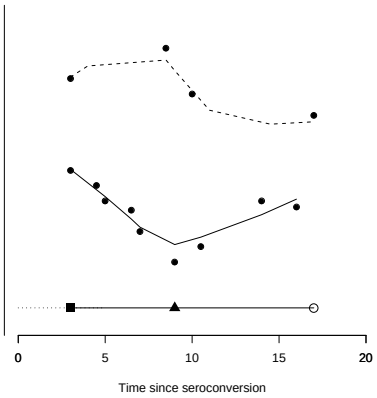

Fig. 1.2. Longitudinal measurements of CD4 cell count and viral load (rhombus) over the time since seroconversion. CD4 cell count pattern (solid line) and viral load pattern (broken line). The time scale is the time since seroconversion: time of entry (square), time of HAART initiation (triangle), time of loss to follow-up (empty point).

time points where markers are not observed, overcoming the problem of missing data and delayed entries, and allows to estimate their trend without error. Furthermore immunological and virological markers' estimates can be used as time-varying predictor variables in an appropriate survival model. Yet models that do not take to account when dropouts occur because of the disease process itself (when they are informative) produce biased estimates of biomarkers' trend. The occurrence of events that cause the end of follow-up could lead to non-ignorable missingness of the biomarkers data, producing overoptimistic estimates if the subjects who leave the study are in poor health or underoptimistic estimates if they are healthy (Touloumi *et al.*, 2002). To include the information held by these dropouts, the extensions to joint modelling of all the processes is required in order to obtain unbiased estimates of unmeasured biomarkers.

Individual characteristics, treatment history and development of opportunistic infections, are carefully recorded for each patient too. The AIDS diagnosis was ascertained through clinical follow-up and through matching with AIDS registries by the original cohort investigators. Only some of the cohorts collected data on AIDS-defining diseases subsequent to the first event, so events after the first are not considered here. The follow-up is artificially censored at the first major modification or at the interruption of the therapy, since the consequent biomarkers' trend may move away from average trend and could be difficult to capture. The

modification and the interruption of the therapy are defined as the change of PI, NNRTI, or ABC, and the suspension of all drugs for more than one week, respectively. Therefore the dropout process can be defined by two different types of event, the naturally occurring censoring of the follow-up due to the study closure or loss to follow-up for noninformative reasons, and the informative withdrawal from the study, i.e. due to the modification and the interruption of therapy. Hence the assumption of independent censoring is unrealistic. Dealing with dependent censoring as competing event for the development of an opportunistic infection allows to obtain unbiased estimates of both the probability of failing from the event of interest and the probability of leaving the study for some cause, when the cause of informative censoring is recognizable.
Furthermore, since after the introduction of HAART, clinical events, such as AIDS or death, have become more rare, researchers' interest is turning towards understanding the possible correlations between biomarkers and further events, such as the suspension and modification of the therapy. What leads an individual to change or interrupt his/her therapy? Is there a common cause underpinning these decisions? Is it possible to identify a particular trend in biomarkers before interruption or change of the therapy? Thus, the need to extend the standard longitudinal models to include competing risks is required in order to deal with the problems generated by dependent censoring and by the interest in modelling the time to multiple endpoints.

1.3 Background and thesis structure

The thesis is organized in two parts, the first focuses on methodological aspects, the second deals with the analyses of the CASCADE data. The last section is set apart for the conclusions and plans for the further research. Since the interest in joint modelling originates from the application, the first part is developed as function of the second.
In the literature, several approaches have been proposed to model CD4 cell count, first separately, then jointly with viral load to understand the natural history of these two biomarkers and their correlation. We will review only some of the models, introduced in the literature since 1980s, restriciting our attention to those which are extensions of the linear mixed effects model (Laird and Ware, 1982). Overall those models can be mainly split into two groups: the first one extends the linear mixed effects model by adding a stochastic process, that depends on the individuals, while the second one models the fixed and random effects using nonparametric functions. Diggle (1988) extends the linear mixed effects model by including a stochastic

process term, namely a Brownian motion, dependent on the ith subject, allowing for high variability in the data. DeGruttola *et al.* (1991) model CD4 cell count by a linear mixed effects model and address the problem of unknown seroconversion date by using an external estimate of the infection time distribution calculated by Bacchetti and Moss (1989). Lange *et al.* (1992) conduct a fully Bayesian analysis of progression of HIV infection using CD4 cell count within a high-dimensional hierarchical model. Their approach allows for individual piecewise linear growth curves with random unobserved change points, unbalanced and incomplete data, several covariates, and unobserved infection times. Galai *et al.* (1993) express the random effects by a stochastic model with damped exponential correlations, including as special cases a first-order autoregressive process and constant autocorrelation. Zeger and Diggle (1994) propose a model with fixed effects, a stationary Gaussian process, and a smooth trend function, estimated by locally adaptive kernel smoothing methods. Taylor *et al.* (1994) use a model that combines the random effects model with a stochastic process allowing correlation between measurements. The stochastic process is the integrated Ornstein-Uhlenbeck (OU) process, which includes Brownian motion and a random effects model as special limiting cases, and it is an underlying continuous-time autoregressive order process for the derivatives of the observations. The motivation for this stochastic process is that the slope of CD4 cell count for an individual can vary, either rapidly or slowly over the time. Kiuchi *et al.* (1995) suggest a piecewise linear growth curve with one random change point for each individual, representing the hypothesized rapid decline of CD4 cell count at time just prior to AIDS diagnosis. Shi *et al.* (1995) modifies the linear mixed effects model, by modelling both the fixed and random effects by cubic B-splines. By using the splines, the adaptation to data may be better but obviously the model's interpretation might be more difficult than by using piecewise linear models.

Sy *et al.* (1997) propose a generalization of the model for univariate longitudinal data (Taylor *et al.*, 1994) to multivariate repeated measures. The model incorporates random effects, correlated stochastic processes, and measurements errors. The stochastic process is the multivariate integrated Ornstein-Uhlenbeck (OU) process. This model allows to investigate the relationship between two disease progression markers by the correlation between the random effects and their serial correlation. Liang *et al.* (2003) study the complicated nonlinear relationship between virologic and immunologic responses by a mixed-effects varying coefficient model with measurement error in covariates. They express the viral load as a function of CD4 cell count and by regression spline method make inference on the parameters. Liang and Zou (2007) propose to model the relation between the two biomarkers by a

semiparametric mixed-effects model. They use the regression spline techniques for inference on model's parameters.

Because of the complexity arising when dealing with joint models in presence of competing events, we will focus on the linear mixed effects models, with the plan to include a stochastic process in the near future. Hence, in the first part of chapter 2 we will introduce the notation and the features of the univariate and bivariate linear mixed effects model. In the second part of the chapter we will describe the survival models in a competing risks framework. In particular, we will provide a definition of competing risks, underlining their pecularities, and we will introduce the extension of the Cox proportional hazards model (Cox and Oakes, 1984) to competing risks (Kalbfleisch and Prentice, 2002) and the Fine & Gray approach (Fine and Gray, 1999) which models the cumulative incidence curves, as opposed to the hazards function.

Having defined these two aspects separately, we will deal with their joint modelling. At the beginning of chapter 3 we will give a brief review of joint modelling from a methodological perspective, highlighting the main approaches proposed in the literature. In particular, we will focus on the model proposed by Elashoff *et al.* (2007), who extend the joint modelling of longitudinal and survival data to competing risks framework. They adopt a linear mixed effects model for the longitudinal measurements and a mixture model (Larson and Dinse, 1985a; Ng and McLachlan, 2003) for the competing risks survival data, and obtain estimates by implementing the EM-based algorithm on the scleroderma lung clinical trial. Then they evaluate if the treatment is effective on at least one of the two endpoints, namely if the treatment can improve %FVC (forced vital capacity) level of a patient or decreases the risk of treatment failure or death. In contrast, motivated by the application, we will propose the joint modelling of a linear mixed effects model and of a Cox proportional hazards model extended to competing risks, fitted by both a frequentist and a Bayesian approach. Specifically, we will describe the steps of the EM-based algorithm and the full conditional distributions of all unknown parameters, after choosing appropriate priors. In order to understand the difference between separate and joint modelling, which sometimes may be tenous also in the results, we need to hightlight the underlying methodological process.

In chapter 4 we will describe the whole dataset and the selection criteria of the patient, who will be included in the analysis. After an explorative data analysis, we will model the CD4 cell count pattern over time since seroconversion and time to three competing events, as a function of CD4 cell count, first separately then jointly for three separate groups of patients. Those gropus are defines by their CD4 cell count at the HAART initiation. Thereafter we will extend the analysis, by mod-

elling the correlation of CD4 cell count and viral load, and the processes leading to each competing events as a function of both biomarkers. Several models have been applied to data provided by the CASCADE collaboration, in particular we cite three papers, in which the joint modelling has been used to answer the question of interest. Pantazis and Touloumi (2005) extend the "joint multivariate random effects" model (Touloumi *et al.*, 1999) to model repeated measures of the two biomarkers simultaneously in presence of informative dropouts due to progression of disease or death and allowing for nonlinear trends. Thiebaut *et al.* (2005) investigate the influence of HIV mode of transmission on virological and immunological response to HAART, by combining a bivariate mixed model for the markers with a lognormal survival model of time-to-dropout using a full parametric approach and implementing an EM-based algorithm. By modelling the survival process, the longitudinal analysis is adjusted for informative dropouts, specifically for exit from the study of patients because of clinical progression, discontinuation of treatment or any potential informative reason that is associated with the latent evolution of the marker. The analyses are developed on time since HAART initiation. Thiebaut *et al.* (2006) study the determinants of immunological and virological response to HAART in HIV patients. For each marker the evolution over time from HAART initiation is studied using a linear mixed effects model. The effects of each potential determinants, i.e. age, gender, year of HAART initiation, elapsed time between seroconversion and HAART initiation, are evaluated on the biomarkers' pattern. The relationship between the two markers is expressed by the covariance matrix of random effects. All analyses are adjusted for potential informative dropout from artificial censoring of follow-up using a joint model for the evolution of markers and the time of censoring. In all three approaches, the aim is to model the biomarkers' pattern, adjusting for informative drop-outs, without a specific interest to model the survival process. On the contrary, we will be mainly interested in modeling time to three competing events, as a function of the biomarkers' patterns. We will address the analysis according to a Bayesian approach, with the aim of implementing an EM algorithm, therefore adopting a frequentist approach, subsequently.
"Although the joint modelling has been proposed several times in the literature, its extension and application change every time according to data and aim of the study".

Part II

Methodology

2

Models and notation

2.1 Longitudinal models

A longitudinal study is defined as a study in which the response for each experimental unit in the study is observed on two or more occasions. In a longitudinal study the main goals to pursue are to characterize patterns subject responses over time, and to investigate the effects of important covariates on these patterns. In order to make it correctly, the analysis of longitudinal data should take into account firstly, the within subject correlation, secondly the measurements taken at unequal time intervals and finally the missing observations. Since the set of observations on each subject tends to be intercorrelated, these correlations must be modelled. Some of the most commonly used within-subject correlation matrices are the independence matrix, when the repeated observations are uncorrelated, the unstructured matrix, when the correlations within any two responses are unknown and need to be estimated, the exchangeable matrix, when the correlation between any two responses of the ith individual is the same, and the autoregressive matrix, when the repeated observations are correlated by an autoregressive process. When each subject is scheduled to be measured at the same set of times, then resulting data is referred as equally-spaced or balanced data, while when subjects are observed at different sets of times and/or there are missing data, then resulting data is referred as a unequally spaced or unbalanced data set. It is very rare to find balanced data sets in longitudinal studies so it is necessary to use some alternative techniques which can handle unbalanced data and missing data.
Diggle *et al.* (2002) review statistical methods for the analysis of discrete and continuous longitudinal data, dealing with different approaches, marginal, transition and random effects models. When a population is of primary interest, fitting marginal models is the most appropriate. In these models, the population-averaged response is modelled as a function of the covariates. The regression coefficients are

interpreted for the population rather than for individuals, so these are known as "population-averaged" models (PA). When the time dependence is central, models for the conditional distribution of y_{ij} given $y_{ij-1}, y_{ij-2}, \ldots$ may be more appropriate, indicating by y_{ij} the measurement at time t_j for ith individual. These are also known as conditional or transition models. The first-order, second-order, third order or higher order autoregressive models belong to this class. Random mixed effects models are more appropriate for the study of an individual's growth. These models are also known as "subject-specific" models (SS).

Let y_{ij} be the response variable and $\boldsymbol{x}_{ij}$ be the vector of explanatory variables observed at time t_{ij} for the subject i, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. The number and the time of measurements may be different for each individual. We denote the mean and the variance of the response variable by $E(y_{ij}) = \mu_{ij}$ and $Var(y_{ij}) = v_{ij}$, respectively. Let $\boldsymbol{y}_i$ the vector of measurements with mean $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ and $n_i \times n_i$ variance-covariance matrix R_i for ith individual. Indicating the complete vector of measurements by $\boldsymbol{y} = (\boldsymbol{y}_1^{'}, \boldsymbol{y}_2^{'}, \ldots, \boldsymbol{y}_m^{'})$, of dimension $N = \sum_{i=1}^{m} n_i \times 1$, the mean of $\boldsymbol{y}$ is $E(\boldsymbol{y}) = \boldsymbol{\mu}$ and the variance matrix $Var(\boldsymbol{y}) = V$.

Under the general model linear, it is assumed that $\boldsymbol{y}$ has a multivariate normal distribution

$$\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{V})$$

and $\boldsymbol{\mu}$ is specified as a linear model given by

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\alpha}$$

where $\boldsymbol{X}$ is the $N \times p$ design matrix and $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown regression coefficients. The specification of V can include three different sources of random variation: random effects, serial correlations and measurement errors. In this case the model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{W}(\boldsymbol{t}) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\theta}, \boldsymbol{W}(t)$ and $\boldsymbol{\epsilon}$ represent random effects, serial correlations, and measurements errors, respectively. $\boldsymbol{Z}$ is a $N \times q$ design matrix with usually $q \leq p$. Together $\boldsymbol{\theta}, \boldsymbol{W}(t)$ and $\boldsymbol{\epsilon}$ have zero mean and specify the variance V. For example, by assuming that the n_i responses on ith individual are indipendent and $\boldsymbol{\theta}_i \sim MVN(\boldsymbol{0}, \Sigma)$, $\boldsymbol{\epsilon_i} \sim N(0, \sigma_\epsilon^2 I_{n_i})$, and $\boldsymbol{W}(t)$ an independent stationary Gaussian processes with mean zero, variance σ_w^2 and correlation function $\rho(u)$ to parameterize furtherly, the covariance matrix can be written as

$$V_i = \boldsymbol{Z}_i \Sigma \boldsymbol{Z}_i^{'} + \sigma_w^2 H_i + \sigma_\epsilon^2 I_i$$

where H_i is the $n_i \times n_i$ simmetric matrix with (j,k)th element $h_{ij} = \rho(|t_{ij} - t_{ik}|)$, and I_{n_i} the $n_i \times n_i$ identity matrix.

According to the specification of V_i, different linear models are considered.
Several estimation methods have been proposed for this model in its general formulation. Laird and Ware (1982), Diggle *et al.* (2002) suggest maximum likelihood (ML) and restricted maximum likelihood (REML) with the remark that REML is usually better than ML. Goldstein (1986) suggest iterative generalized linear model (IGLS) and restricted IGLS (RIGLS) for more general multilevel structure. Bates and Pinheiro (1998) propose EM estimation followed by Newton-Rhapson or quasi-Newton optimization of the loglikelihood or the log-restricted-likelihood. Zeger and Karim (1991) formulate a Bayesian method using Gibbs sampling.
For a more general longitudinal model with non-Gaussian outcome, Zeger and Liang (1986) propose an extension of the generalized linear model (GLM). 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. In this model the mean of $\boldsymbol{y}_i$ is modelled by

$$\boldsymbol{\mu}_i = h(\boldsymbol{X}_i\boldsymbol{\alpha}).$$

By an iterative procedure, a consistent estimator of $\boldsymbol{\alpha}$ is obtained, by solving the generalized estimating equation. This approach is an example of the population averaged model (PA) (Zeger *et al.*, 1988)
Generalized linear mixed model (GLMM) is an extension of GLM by including random effects, or more general multilevel or hierarchical structure in the model. This approach models the mean of $\boldsymbol{y}_i$ conditional on random effects, that is

$$E(\boldsymbol{y}_i|\boldsymbol{\theta}) = h(\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z_i}\boldsymbol{\theta_i}).$$

This model is also known as subject specific model (SS) (Zeger *et al.*, 1988).
Specifically in the next two sections we will introduce the linear mixed effects model (Laird and Ware, 1982), by an univariate and bivariate approach respectively.

2.1.1 Univariate linear mixed effects model

Laird and Ware (1982) defined a family of models that include both growth models and repeated-measure models as special cases. Both models belong to class of two-stage models, the first one explains the within-subject variation by the natural development (Potthoff and Roy, 1964; Rao, 1965; Fearn, 1975; Ware, 1983), while the second one typically assumes constant individual effects over the time (Hayes, 1973). By using the same therminology adopted above, the model is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

where $\boldsymbol{\alpha}$ denotes a $p \times 1$ vector of unknown fixed effects, $\boldsymbol{X}_i$ is a known $n_i \times p$ design matrix linking $\boldsymbol{\alpha}$ to set of longitudinal measurements $\boldsymbol{y_i}$, $\boldsymbol{\theta}_i$ denotes a $q \times 1$ vector of unobservable random effects, with $q \leq p$, $\boldsymbol{z}_i$ is a known $n_i \times q$ design matrix linking $\boldsymbol{\theta}_i$ to $\boldsymbol{y_i}$, $\boldsymbol{\epsilon}_i$ is a within-individual residuals vector. Furthermore $\boldsymbol{\epsilon}_i$ are assumed to be indipendent and normally distributed with mean $\mathbf{0}$ and $n_i \times n_i$ positive-defined covariance matrix Σ_ϵ. Σ_ϵ depends on subject i by its dimensionality but not by the parameters which form it. At the first stage the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}_i$ are considered fixed-effects. At the second stage only the parameters $\boldsymbol{\alpha}$ are treated as fixed-effect, while the parameters $\boldsymbol{\theta}_i$ are normally distributed with mean $\mathbf{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{\Sigma}$. The $\boldsymbol{\theta}_i$ are distributed indipendently of each other and of the within-subjects residuals $\boldsymbol{\epsilon_i}$.
Marginally the vector $\boldsymbol{y}_i$ is normally distributed with mean $\boldsymbol{x}_i\boldsymbol{\alpha}$ and variance-covariance matrix $\Sigma_\epsilon + \boldsymbol{Z}_i\Sigma\boldsymbol{Z}_i^{'}$. Under this model $\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{\theta}_i$ can be thought of as the true value of repsonse variable over time and the correlation between the repeated meaurements on an individual arises from an individual's deviation from overall effect. When $\Sigma_\epsilon = \sigma_\epsilon^2 I_{n_i}$, where I_{n_i} denotes the $n_i \times n_i$ identity matrix, the n_i responses of ith individual are independent, conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\theta_i}$. Ordinary iterative methods for maximising likelihoods, such as newthon Raphson, the Fisher scoring and the EM algorithm, are used to obtain maximum likelihood or restricted maximum likelihod estimates for unknown parameters (Dempster *et al.*, 1977; Dempster *et al.*, 1981; Harville, 1977; Laird, 1982).

2.1.2 Bivariate linear mixed effects model

Many situations arise in which two or more response variables are observed on each individual, simultaneously or not. If the correlation between the two variables was high, the effect of covariates on response variables patterns could be estimated more efficiently compared to that estimated separately for each response variable. A simultaneous modelling approach for bivariate response repeated-measures data requires a generalization of usual mixed effects models for a single response variable. A possible approach to model the dependence between the two variables is by the random effects. Let $\{y_{ij}^1 : j = 1, \ldots, n_i^1\}$ and $\{y_{ij}^2 : j = 1, \ldots, n_i^2\}$ be two sets of longitudinal quantitative measurements at times $\{t_{i1}^1, t_{i2}^1, \ldots, t_{in_i^1}^1\}$ and $\{t_{i1}^2, t_{i2}^2, \ldots, t_{in_i^2}^2\}$ respectively for ith individual, $i = 1, \ldots, m$. Hence the bivariate mixed effects models for the vector $(\boldsymbol{y}_i^{'}, \boldsymbol{y}_i^{''})$ is given by

$$\begin{pmatrix} \boldsymbol{y}_i^1 \\ \boldsymbol{y}_i^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_i^1 & 0 \\ 0 & \boldsymbol{X}_i^2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\alpha}^1 \\ \boldsymbol{\alpha}^2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{Z}_i^1 & 0 \\ 0 & \boldsymbol{Z}_i^2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\theta}_i^1 \\ \boldsymbol{\theta}_i^2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_i^1 \\ \boldsymbol{\epsilon}_i^2 \end{pmatrix} \qquad (2.2)$$

where $\boldsymbol{\alpha}^k$ denote a $p^k \times 1$ vector of unknown fixed effects, $\boldsymbol{X}_i^k$ be a known $n_i^k \times p^k$ design matrix respectively linking $\boldsymbol{\alpha}^k$ to set of longitudinal measurements $\boldsymbol{y}_i^k$, $\boldsymbol{\theta}_i^k$

denote a $q^k \times 1$ vector of unobservable random effects, with $q^k \leq p^k$, $\boldsymbol{Z}_i^k$ be a known $n_i^k \times q^k$ design matrix linking $\boldsymbol{\theta}_i^k$ to $\boldsymbol{y}_i^k$, and finally $\boldsymbol{\epsilon}_i^k$ be a within-individual residuals vector, $k = 1, 2$. Let $\boldsymbol{\theta}_i$ be the vector $(\boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2)$, and $\boldsymbol{\epsilon}_i$ be the vector $(\boldsymbol{\epsilon}_i^1, \boldsymbol{\epsilon}_i^2)$. Hence

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \Sigma_\epsilon) \qquad \boldsymbol{\theta}_i \sim N(\mathbf{0}, \Sigma)$$

where

$$\Sigma_\epsilon = \begin{pmatrix} \Sigma_{\epsilon 1} & 0 \\ 0 & \Sigma_{\epsilon 2} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}.$$

Σ_{ϵ_1} and Σ_{ϵ_2} are a $n_i^1 \times n_i^1$ and $n_i^2 \times n_i^2$ covariance matrices for $\boldsymbol{y}_i^1$ and $\boldsymbol{y}_i^2$ respectively. The covariance matrix of random effects Σ is partitioned in four sub-matrices, Σ_1 being the covariance matrix including variance and covariance of random effects of the first response variable, Σ_2 being the covariance matrix including variance and covariance of random effects of the second response variable, and $\Sigma_{12} = \Sigma_{21}$ being the matrix of covariance between random effects of each response variable. The correlation between the two response variables is taken in account by Σ_{12}.

2.2 Survival models

Survival data arise when the aim is to study the time elapsed from some particular starting point to the occurrence of an event. In clinical studies the starting point of the observation is usually a medical interventation or the beginning of a treatment study. In epidemiological studies the starting point may be the birth or the beginning of an exposure to some risk factor. The terminal event may be death or a prespecified event of interest. Survival analysis is useful whenever the researcher is interested not only in the frequency of occurrence of a certain type of event, but also in the time process underlying such occurrence (Cox and Oakes, 1984; Fleming and Arrington, 1991; Andersen *et al.*, 1993; Marubini and Valsecchi, 2004; Kalbfleisch and Prentice, 2002).

Let the non-negative random variable T denote the time until the occurrence of the event of interest. Primary interest in survival analysis lies in estimation and testing regarding the distribution of T. The probability distribution of T can be specified in many ways, three of which are particurarly useful in survival applications, the survivor function, the probability density function, and the hazard function, respectively given by

$$S(t) = P(T > t)$$

$$f(t) = \lim_{\Delta t \to 0} P(t \leq T < t + \Delta t)$$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} log S(t)$$

A distinctive characteristic of survival data is that the event of interest may not be observed on every esperimental unit. This feature is known as censoring. Censoring can arise because of time limits and other restrinctions depending on the nature of the study. In clinical and epidemiological studies, censoring can be "produced" by different causes, such as predeterminate duration of the study or causes that seldom can be determinate. In the first case for example, ethical, scientific and economic reasons could suggest that the study continues until a prespecified time point and then the time to event of interest is known precisely only on those subjects who present the event before that time point. For the remaining subjects it is only known that the time to the event is greater than the observation time. In the second case, some subjects may be unwilling or unable, for some reasons, to continue partecipating in the study and providing follow-up information. These subiects are called "dropouts". In both cases, those incomplete data are considered right censored.

Let C denote the censoring time. Then (T, C) are latent data, while (U, Δ) are observed data, where $U = min(T, C)$, $\Delta = I(T \leq C)$ and $I(\cdot)$ is the indicator function. While the distribution funcion $S(t)$ can be consistently estimated when the data are uncensored, neither $\lambda(t)$ nor $S(t)$ is identificable or consistently estimable if one observes (U, Δ) (Fleming and Lin, 2000). Observing (U, Δ) rather than T for all partecipants only allows one to consistently estimate

$$\lambda^{\dagger}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t, C > t)}{\Delta t}$$

Therefore, in most survival applications, a key assumption is made regarding the following equality

$$\lambda(t) = \lambda^{\dagger}(t)$$

for all t such that $P(U > t) > 0$. A sufficient condition for the validity of this assumption is the indipendence of T and C. The class of censoring mechanisms, which satisfies this condition, is called "independent censoring". For example, the first-type censoring (predeterminate duration of the study) or second-type censoring (the study continues until the dth smallest failure time occurs, at which time all surviving items are censored) are independent censoring. What it requires is that, at time t, study items cannot be censored because they appear to be at unusually high or low risk of failure. In such setting $\lambda(t) << \lambda^{\dagger}(t)$ and then we would overstimate the true $S(t)$. The methods introduced in the next sections are based on the assumption of independent censoring.

2.2.1 Competing risks

A competing risk is an event whose occurrence either precludes the occurrence of another events under examination or alters the probability of occurrence of these other events (Crowder, 2001; Putter *et al.*, 2007). In the first case, for example in mortality studies, competing risks analysis tackles the problem of how an increase or decrease of one cause of death impacts on the risks of dying from other causes. Analyses of this kind allow the evaluation of whether an excess in the probability of dying from one cause, for example stroke, can be partially attributed to the deficit from other causes, such as infarction. In the second case, for example in cancer studies, competing risks usually include relapse of the cancer and death in remission. Here the interplay between the competing risks of relapse and death in remission gives the complete story about the efficacy of the treatment. In this setting the time to the first failure of any type would appear to be the most clinically relevant endpoint to the patient, i.e. disease-free survival defined as the time to disease recurrence or death.
Suppose an individual is exposed to K types of events which cannot occur simultaneously. Under the competing risks framework only the first of these event is considered. Assuming the terminology adopted by Tai *et al.* (2001), we define the following quantities:
$L \in 1, 2, \ldots, K$ the type of the first event

T_l the time to event of type l, $l \in (1, \ldots, K)$

$\lambda_l^*(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_l < t + \Delta t | T_l > t)}{\Delta t}$ the hazard function for event-type l

$S_l(t) = P(T_l > t)$ the survivor function for event-type l

$f_l(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T_l < t + \Delta t)}{\Delta t}$ the unconditional probability function of T_l

$I_l(t) = P(T \leq t \wedge L = l)$ the cumulative incidence function for event-type l

$T = \min(T_1, T_2, \ldots, T_K)$ the time to first event

$S(t) = P(T > t)$ the event-free survivor function

$\lambda_l(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T_l < t + \Delta t | T > t)}{\Delta t}$ the hazard function for event type l, conditional

on no other event having occurred.
The potential survival time T_l are contrasted with the observed survival time T. The latent failure time approach focuses on the joint distribution of the times to

K different events, as described by the joint survival function, $P(T_1 > t_1, T_2 > t_2, \ldots, T_K > t_k)$. Without additional assumptions, the joint survival function is not identifiable from the observed data (**?**). Hence the marginal probabilities, $S_l(t)$, are not identifiable too, unless the times to competing events are independent.
Since in a competing risks framework we are interested in the first event that occurs in every individual, we focus on the hazard function for failure type l, $\lambda_l(t)$. However, this measure, being an instantaneous risk function, does not quantify the cumulative probability of developing a specific event or, for example in a clinical trial, the ultimate benefit of a treatment to the patient (it could happen that hazard functions for two treatments cross and in this case one is not able to say which treatment leads to the smaller chance of event of interest). A good alternative is the cumulative incidence curve, which estimates the marginal cumulative probability of a particular event occurring as first one. This curve has a straighforward interpretation. In the first example it is, for each event type and at any given time, the cumulative probability of dying for stroke and the cumulative probability of dying for infarction, respectively . In the second example it is, respectively for each event type and at any given time, the cumulative probability of relapse of cancer as first event and the cumulative probability of death in remission as first event. Note that, the marginal cumulative probability of relapse will be low if the marginal cumulative probability of death in remission is high and viceversa, since the sum of these quantities can not exceed 1. Hence it is advisable to consider all marginal probability curves simultaneously in order to interpret them appropriately.
In competing risks framework different approaches have been proposed, parametric and non-parametric, some of them proper to estimate the hazard function (Cox, 1972; Gaynor *et al.*, 1993; Lunn and McNeil, 1995), other the cumulative incidence curves (Fine, 2001; Klein and Andersen, 2005, Jeong and Fine, 2006).
Kalbfleisch and Prentice (2002) propose an estimate of $\lambda_l(t)$, $S_l(t)$ and $I_l(t)$, obtained by maximizing the likelihood function in a competing risks framework. It is not based on any strong assumption, as the independence of competing risks, other than the usual independent and non informative censoring mechanism.
Suppose n subjects under study give rise to data $(t_i, \delta_i, l_i, \boldsymbol{w}_i)$, $i = 1, \ldots, m$, where $\delta_i = 0$ if the ith subject is censored, $\delta_i = 1$ if the ith is failed, t_i is the censoring time ($\delta_i = 0$) or the failure time ($\delta_i = 1$), l_i is the type of failure and $\boldsymbol{w}_i$ is the row vector of s covariates associated with the ith individual. The reasonable adopted convention is that censored times follow failures in case that recorded times coincide. The likelihood function is proportional to

$$\prod_{i=1}^{n} \left[f_{l_i}(t_i; \boldsymbol{w}_i)^{\delta_i} S(t_i; \boldsymbol{w}_i)^{1-\delta_i} \right] =$$

$$= \prod_{i=1}^{n} \left[\lambda_{l_i}(t_i; \boldsymbol{w}_i)^{\delta_i} S(t_i; \boldsymbol{w}_i)\right] = \prod_{i=1}^{n} \left\{\lambda_{l_i}(t_i; \boldsymbol{w}_i)^{\delta_i} \prod_{l=1}^{K} exp\left[-\int_0^{t_i} \lambda_l(u; \boldsymbol{w}_i) du\right]\right\}$$

Upon rearragement the likelihood factor for the lth failure type is precisely that which would be obtained by regarding all failures of type different from l as censored at the individual's failure time. The "non-parametric" estimation technique of Kaplan-Meier (Kaplan and Meier, 1958) can be generalized to include competing risks. Let $t_1 < t_2 < \ldots < t_r$ denote the r failure times for failures of type l, $l = 1, \ldots, K$ and suppose failure type l occurs with multiplicity d_{lj} at time t_j. The contribute to likelihood of failures type l is given by:

$$\prod_{j=1}^{r} \left\{\left[S_l(t_j^-) - S_l(t_j)\right] \prod_{q=1:q\neq l}^{K} S_q(t_j^-)\right\}^{d_{lj}} \prod_{w} S_l(t_{jw})$$

where t_{jw} denote the censored times between t_j and the next failure time, the probability of failure at t_j is $\left[S_l(t_j)^- - S_l(t_j)\right]$, where $S_l(t_j^-) = \lim_{x\to 0} S_l(t_j - x)$, and $S_l(t_j)$ is the contribution to the likelihood of a survival time censored at t_j so that the observed censoring time t_j suggests only that the unobserved failure time is greater than t_j. The maximum likelihood estimates $\widehat{S}_l(t)$ is a generalization of the usual concept used in the parametric models.

The likelihood function is given by the product of single components for each failure type. It follows that the "non-parametric" maximum likelihood estimator of $S_l(t)$ is given by

$$\widehat{S}_l(t) = \prod_{j:t_j \leq t} \left(\frac{n_j - d_{lj}}{n_j}\right)$$

where d_{lj} is the number of failure type l at t_j and n_j is the number of subjects under study at risk just prior to t_j. The corresponding estimator of the lth cause-specific hazard function is

$$\widehat{\lambda}_l(t_j) = \frac{d_{lj}}{n_j}$$

and the estimator of the cause-specific cumulative hazard function, also known as Nelson-Aalen estimator,

$$\widehat{H}_l(t) = \sum_{j:t_j \leq t} \frac{d_{lj}}{n_j}$$

Note that under assumption of independence between competing risks

$$\widehat{S}(t) = \prod_{l=1}^{k} \widehat{S}_l(t)$$

is the overall Kaplan-Meier survivor function estimator. The cause-specific cumulative incidence function for failure of type l is estimated by

$$\widehat{I}_l(t_j) = \sum_{j:t_j \leq t} \widehat{S}(t_{j-1})\widehat{\lambda}_l(t_j)$$

where $\widehat{S}(t_{j-1})$ is the Kaplan-Maier estimate of the event-free survival function, that is, considering failure of any kind. This definition implies that the cumulative incidence is a function of the hazards of all the competing events and not solely of the hazard of the event to which it refers (Coviello and Boggess, 2004). Indeed the incidence rates are computed by weighting the hazard of first failure, $\widehat{\lambda}_l(t_j)$, with the event-free survival estimates in the preceding time. This method does not make any assumption about independence of competing risks and lets a subjects "fail" only once. The sum of all cumulative incidences, given by

$$\widehat{I}(t) = \sum_{l=1}^{K} \widehat{I}_l(t) = 1 - \widehat{S}(t)$$

equals the complement of the overall Kaplan-Meier estimate of survival considering failures of any kind.

2.2.2 Cox proportional hazards model

One of the most used model for identifying differences in survival due to treatment and prognostic factors in clinical trials and for studying the effect of exposure allowing for confounders in cohort studies is the proportional hazards model (Cox, 1972). By extending the Cox model to the competing risks, the cause-specific hazard function for a failure of type l, is given by

$$\lambda_l(t_j; \boldsymbol{w}) = \lambda_{0l}(t_j) exp(\boldsymbol{w}\boldsymbol{\phi}_l)$$

$l = 1, \ldots, k$ where $\lambda_{0l}(t_j)$ is the arbitrary unspecified baseline hazard function for T_l and $\boldsymbol{\phi}_l$ is the vector of unknown parameters. Both the underlying hazard $\lambda_{0l}(t_j) \geq 0$ and the vector of regression coefficients are specific to each of the K failure types. The hazard depends on both time and covariates, but through two separate factors: the first, $\lambda_{0l}(t)$, is a function of time and the failure type, it is left arbitrary but is assumed to be the same for all subjects; the second is a quantity which depends on the failure type and the individual covariates through the vector $\boldsymbol{\phi}_l$ of regression coefficients. The proportional hazards model is not a fully non-parametric model since it does not specify the form of the baseline hazard function, $\lambda_{0l}(t)$, but it does, however, specify the hazard ratio for two indviduals with covariate vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$ and for this reason it is defined as a semiparametric model. In fact the hazard ratio for failure type l is given by

$$\frac{\lambda_l(t; \boldsymbol{w}_1)}{\lambda_l(t; \boldsymbol{w}_2)} = \frac{\lambda_{0l}(t) exp(\boldsymbol{w}_1\boldsymbol{\alpha})}{\lambda_{0l}(t) exp(\boldsymbol{w}_2\boldsymbol{\alpha})} = exp[(\boldsymbol{w}_1 - \boldsymbol{w}_2)\boldsymbol{\alpha}]$$

Hence this model is a proportional hazard regression model for each failure type, since it assumes that the failure rates of any two individuals are proportional, given that the ratio does not depend on time. The second assumption underlying this model, is that the vector of covariates $\boldsymbol{w}$ acts in a multiplicative way on the hazard function or, equivalently, in an additive way on the logarithm of the hazard function. Let $t_{l1}, t_{l2}, \ldots, t_{lr_l}$ denote the r_l times of failure of type l event, $l = 1, \ldots, K$, and $\boldsymbol{w}_{lj}$ the regression vector for individual that fails at t_{lj}. The partial likelihood for $\boldsymbol{\phi_l}$, $l = 1, \ldots, K$ is given by:

$$L(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K) = \prod_{l=1}^{K} \prod_{j=1}^{r_l} \frac{exp(\boldsymbol{w}_{lj}\phi_l)}{\sum_{i \in R(t_{lj})} exp(\boldsymbol{w}_i \boldsymbol{\phi}_l)}$$

Upon rearragement the likelihood factors into a separate components for each failure type $l = 1, \ldots, K$. Hence, estimation of $\boldsymbol{\phi}_l$'s can be conducted by applying standard asymptotic likelihood techniques individually for j factors. The likelihood factor for the lth failure type is precisely that which would be obtained by regarding all failures of type different from l as censored at the individual's failure time. The estimates of the underlying hazard $\widehat{\lambda}_{0l}(t_j)$ is computed by assuming it to be zero except at times at which a failure of type l occurs, in that case:

$$\widehat{\lambda}_{0l}(t_j) = \frac{d_{lj}}{\sum_{i \in R(t_j)} exp(\boldsymbol{w}_i \widehat{\boldsymbol{\phi}}_l)}$$

where $R(t_j)$ is the set of individuals at risk just prior to t_j and d_{lj} is the number of failure of type l at time t_j. Once obtained the maximum likelihood estimators $\widehat{\boldsymbol{\phi}}_1, \widehat{\boldsymbol{\phi}}_2, \ldots, \widehat{\boldsymbol{\phi}}_K$ and the estimator $\widehat{\lambda}^*_{0l}(t_j)$, it is possible to estimate the functions $S_l(t; \boldsymbol{w})$, and the corresponding estimator of the cumulative incidence function

$$\widehat{I}_l(t; \boldsymbol{w}) = \sum_{j: t_j \leq t} \widehat{S}(t_{j-1}; \boldsymbol{w}) \widehat{\lambda}_l(t_j; \boldsymbol{w})$$

$l = 1, \ldots, K$, where $\widehat{S}(t_{j-1}; \boldsymbol{w})$ is the Kaplan-Meier estimate of the overall survival function and $\widehat{\lambda}_l(t_j; \boldsymbol{w}) = d_{lj}/n_j$ is the estimate of the hazard of l failure type.
The proportional hazards model requires that, for any two covariates sets $\boldsymbol{w}_1, \boldsymbol{w}_2$, $\lambda_l(t; \boldsymbol{w}_1) \propto \lambda_l(t; \boldsymbol{w}_2)$, $l = 1, \ldots, K$. Although this relation is descriptive of many situations, there are important factors, whose different levels produce hazards function, which differ from proportionality. Suppose there is a factor that occurs on q levels and for which the assumption of proportionality is violated. The hazard function for a failure type l in the jth stratum of this factor is given by

$$\lambda_{lj}(t; \boldsymbol{w}) = \lambda_{0lj}(t) exp(\boldsymbol{w}\boldsymbol{\phi}_l)$$

$j = 1, \ldots, q$, where $\boldsymbol{\phi}_l$ and $\boldsymbol{w}$ are invariant for every j and $\lambda_{0l1}, \lambda_{0l2}, \ldots, \lambda_{0lq}$ are allowed to be arbitrary and completely unrelated. The likelihood is given by

$$L(\boldsymbol{\phi}_l) = \prod_{j=1}^{q} L_j(\boldsymbol{\phi}_l)$$

where $L_j(\boldsymbol{\phi}_l)$ is the marginal likelihood of $\boldsymbol{\phi}_l$ arising from the jth stratum alone. Once an estimates of β is obtained it is possible to give estimates of the survivor functions in each of the q strata separately. This provides a graphical check of the appropriateness of a proportional hazards modelling for these factors used in defining strata.
Both model, the proportional hazards model and the proportional hazards model stratified, can be extended in order to include time-varying covariates.

2.2.3 Cumulative incidence curve

The standard analysis for competing risks data involves modeling and analyzing the effect of factors on tha cause-specific hazard function $\lambda_l(t)$ for each $l = 1, \ldots, K$. Yet, the cause-specific hazard function does not have direct interpretation in terms of survival probabilities for a particular failure type. Indeed the effect of a factor on the cause-specific hazard for a particular failure type could be quite different from its effect on the corresponding cumulative incidence function (Gray, 1988). It is important consider this second aspect too. In the previous section we have seen that it is possible to predict the incidence cumulative function for an individual with certain covariates by combining the estimates of the cause-specific hazard functions from the partial likelihood approach. However these procedures do not allow to directly assess the effect of a covariate on the cumulative incidence function. Hence, Fine and Gray (1999) propose a semiparametric model for the cumulative incidence function for event of interest, $l = 1$ conditional on the covariates, $I_1(t; \boldsymbol{w}) = P(T \leq t, L = 1 | \boldsymbol{w})$. Assume that for some unknown increasing function $g(\cdot)$,

$$g[I_1(t; \boldsymbol{w})] = \lambda_0(t) + \boldsymbol{w}\boldsymbol{\phi}_1$$

where $\lambda_0(t)$ is an unspecified, invertible and monotone increasing function and $\boldsymbol{\phi}_1$ is a $s \times 1$ parameter vector. For two individuals with two covariate vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$, the cumulative incidence functions satisfy a vertical shift model after trasformation, $g[I_1(t; \boldsymbol{w})] - g[I_1(t; \boldsymbol{w})] = (\boldsymbol{w}_1 - \boldsymbol{w}_2)\phi_1$ for each t. On the scale of $g(\cdot)$ the regression coefficients are a measure of distance from the baseline marginal probability fuction, $g^{-1}[\lambda_0(t)]$, for which the covariates are identically 0. Assume

$$g(u) = log[-log(1 - u)]$$

and consider the subdistribution hazard, as defined (Gray, 1988)

$$\lambda_1^{'}(t;\boldsymbol{w}) = lim_{\Delta t\to 0}\frac{P(t<T\leq T+\Delta t, L=1|T>t\cup(T\leq t\cap L\neq 1),\boldsymbol{w})}{\Delta t} =$$

$$= lim_{\Delta t\to 0}\frac{1}{\Delta t}\frac{P(t<T\leq T+\Delta t, L=1)}{P(T>t\cup(T\leq t\cap L\neq 1),\boldsymbol{w})} =$$

$$= lim_{\Delta t\to 0}\frac{1}{\Delta t}\frac{P(t<T\leq T+\Delta t, L=1}{1-P((T>t\cap L=1),\boldsymbol{w})} =$$

$$= \frac{d}{dt}\Big(\frac{I_1(t;\boldsymbol{w})}{1-I_1(t;\boldsymbol{w})}\Big) = -\frac{d}{dt}\big(-log[1-I_1(t;\boldsymbol{w})]\big)$$

In the previous sections we have considered the hazard function so defined

$$\lambda_1(t;\boldsymbol{w}) = lim_{\Delta t\to 0}\frac{P(t<T\leq t+\Delta t, L=1|T>t,\boldsymbol{w})}{\Delta t} =$$

$$= \frac{d}{dt}\Big(\frac{I_1(t;\boldsymbol{w})}{S(t)}\Big) = \frac{d}{dt}\Big(\frac{I_1(t;\boldsymbol{w})}{1-\sum_{l=1}^{k}I_l(t;\boldsymbol{w})}\Big)$$

One can think of $\lambda_1^{'}(t;\boldsymbol{w})$ as the hazard function for the improper variable $T^* = I(L=1)\times T + [1-I(L=1)]\times\infty$ and T^* has distribution function equal to $I_1(t;\boldsymbol{w})$ for $t<\infty$ and $P(T^*=1|\boldsymbol{w}) = P(T<\infty, L\neq 1|\boldsymbol{w}) = 1-I_1(\infty;\boldsymbol{w})$ for $t=\infty$. Although the risk set associated with the subditribution hazard $\lambda_1^{'}$ is innatural, since in reality those individuals who have already failed from causes other than $L=1$ prior to time t are not at "risk" at t, it is important to consider the possible dependence between failure types. Indeed considering the proportional hazard model $\lambda_l(t;\boldsymbol{w}) = \lambda_0(t)exp(\boldsymbol{w}\boldsymbol{\phi}_l)$, $\boldsymbol{\phi}_l$ represents the effect of covariate vector on hazard l. The cause-specific hazard for each failure type at any timepoint, t, is the instantaneous risk of developing that failure as first event, conditional on being alive and event-free just prior to t. This conditioning means that the cause-specific hazard can not be truly specific to the event of interest because factors which directly influence other failure types can have an indirect effect on the event of interest. Since, in the cause-specific model, individuals who fail for a different failure type first are censored at this time, the estimates $\boldsymbol{\phi}_l$ can be interpreted as the effect of factors on a specific event in the absence of all other events only by assuming that events occurr independently of each other. In that case we could obtain the equality between cause-specific hazard and subdistribution hazard. Yet the independence is an unverifiable and unrealistic assumption making the cause-specific model difficult to interpret. Under a proportional hazards specification with $\lambda_1^{'}(t;\boldsymbol{w}) = \lambda_{01}^{'}(t)exp(\boldsymbol{w}\boldsymbol{\phi}_1)$ where $\lambda_{01}^{'}$ is a completely unspecified, non-negative function we obtain

$$log[1 - I_1(t; \boldsymbol{w})] = -\int_0^t \lambda_1^{'}(u; \boldsymbol{w})du = -\int_0^t \lambda_{01}^{'}(u)exp(\boldsymbol{w}\boldsymbol{\phi}_1)du$$

$$log[-log(1 - I_1(t; \boldsymbol{w}))] = log\Big[\int_0^t \lambda_{01}^{'}(u)exp(\boldsymbol{w}\boldsymbol{\phi}_1)du\Big] = \boldsymbol{w}\boldsymbol{\phi}_1 log\Big[\int_0^t \lambda_{01}^{'}(u)du\Big]$$

Thus the regression coefficient and baseline hazard from the Cox model has a straightforward interpretation that does not depend on the probabilitic structure of the subdistribution hazard. The cumulative incidence function for failure 1 is given by

$$I_1(t; \boldsymbol{w}) = 1 - exp\Big(-\int_0^t \lambda_{01}^{'}(u)exp(\boldsymbol{w}\boldsymbol{\phi}_1)du\Big)$$

The main advantage of the subdistribution methodology is that through simple testing, model selection and prediction procedures it is possible to see the direct effect of each covariate on the cumulative incidence curves. Using the partial likelihood principle and weighting techniques, Fine and Gray derive estimation and inference procedures for the finite-dimensional regession parameters under a variety of censoring scenario (complete data without censoring, censoring complete data and incomplete data). They give an uniformly consistent estimator for the predicted cumulative incidence for an individual with certain covariates and confidence intervals and bands can be obtained analitically or with an easy-to-implement simulation technique.

As an example in the real context, we consider the survival times of 506 patients with prostate cancer who are randomly allocated to a treatment with diethylstilbestrol (Lunn and McNeil, 1995). We estimate the effect of treatment on the hazard to fail from cancer, cardio-vascular disease and by other causes respectively, by the proportional hazards models. Furthermore we evaluate the effect of therapy on the cumulative incidence functions directly, by the Fine & Gray model. The results are reported in table 2.1.

Table 2.1. Estimated effect of therapy on hazard to fail from competing events and on the their cumulative incidence functions

Effect of treatment	**Proportional hazards**		**Fine & Gray**	
Cancer	-0.391	(-0.760, -0.761)	-0.412	(-0.732, -0.092)
CDV	0.169	(-0.162, 0.502)	0.269	(-0.063, 0.601)
Other	-0.456	(-0.970,0.063)	-0.407	(-0.923, 0.109)

Finally we represent the complement of Kaplan Meier function, the nonparametric cumulative incidence functions and the cumulative incidence functions estimated by Fine & Gray model for three competing risks, cancer, cardiovascular disease

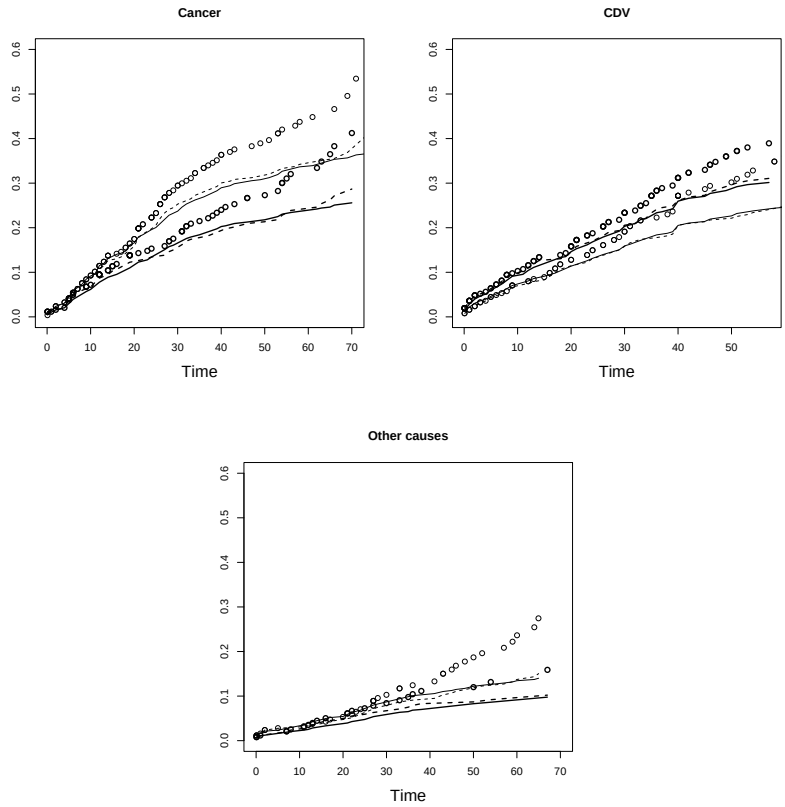

Fig. 2.1. Complement of Kaplan-Meier function (points), nonparametric cumulative incidence function (broken line), Fine & Gray cumulative incidence function (solid line) for treated (thick line) and not treated (thin line) patients.

and other diseases respectively, in figure 2.1. It is clear that the complement of the Kaplan-Meier overestimates the true failure probability (Arriagada *et al.*, 1992; Gooley *et al.*, 1999). The complement of Kaplan-Meier function, given by

$$\widehat{I}_l(t) = 1 - \widehat{S}_l(t)$$

is interpreted as the cumulative probability of failure by time t for event-type l only if the risk of failure for other causes could be removed. This is a predictive probability for an hypothetical setting (Pepe and Mori, 1993). Furthermore, since an individual can experience more than one event-type, this method is no longer valid in a framework of competing risks, in which is considered only the first failure of every subject. So the global cumulative incidence function given by

$$\widehat{I}(t) = \sum_{l} \widehat{I}_l(t)$$

may exceed the total probability of failure, $P(T \leq t)$.

3

Joint modelling of longitudinal and survival data

3.1 Overview of joint modelling of longitudinal and survival data

The most studies in the medical-epidemiological field are characterized by both covariates which vary with time and the time to event of interest, i.e. death or a disease. Usually the covariates are measured intermittently, at varying time points and in unequal numbers for different study participants, and may be prone to measurements error because of laboratory and/or physiological variations. At the same time these measurements may be important predictors of survival. Tsiatis *et al.* (1995) propose an approach developed in two stages dealing with survival as a function of the longitudinal covariate's measurements, where in the first stage the covariate is modelled by growth curve model with random effects, and in the second stage the modelled value is simply plugged into the partial likelihood for the Cox's model with time-dependent covariates, and the partial likelihood is then maximized. This approach is computationally straightforward, it allows for an easy analysis of the data with existing software packages and it reduces the bias in a model with time-dependent covariates measured with error. Yet, since the two-stages model does not use the information provided by survival process in modelling the longitudinal one, the data are not used so efficiently as they could be, and mainly the informative loss to follow-up could generate biased covariate's estimate if it is not considered. Hence the need to model the longitudinal and the survival process jointly. Estimating the parameters that describe the covariate process and those that describe the time-to-event as a function of covariate process simultaneously allows to use not only the observed covariates to predict the survival but also the survival information to model the true covariate process over the time. A joint model is comprised of two linked submodels, one for the "true" longitudinal process $Y_i(t)$ and one for the failure time T_i, where i denotes the ith individual, along with

additional specifications and assumptions that allow a full representation of the joint distribution of the observed data $(Y_i(t), T_i, \Delta_i)$. The longitudinal and disease process are assumed to be independent across subject i. The joint likelihood can be expressed as the product of two density functions $f(T|Y)f(Y)$ or $f(Y|T)f(T)$, according to aim of the study (Hogan and Laird, 1997). If the primary outcome of interest is the time to event and the longitudinal measurements may be predictive of survival, the first model is used, otherwise if the main objective is to characterize changes over time in the longitudinal data process accounting for loss to follow-up by the survival process (Wu and Carroll, 1988), the second model is used. Several methods have been proposed to model the longitudinal and survival data jointly, by both frequentist and Bayesian approaches. We will discuss some of them, first by a frequentist point of view, then by a Bayesian one.
DeGruttola and Tu (1994) jointly model disease progression and failure times using a longitudinal data model given by a random effects model

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{z}_i\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i$$

and a survival model, given by

$$t_i = w_i^{'}\xi + \lambda^{'}\theta_i + r_i$$

where $\boldsymbol{y}_i$ is a $n_i \times 1$ vector of repeated measurements on the ith subject, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are known design matrices, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown fixed effects, $\boldsymbol{\theta}_i$ is a $q \times 1$ vector of unobservable random effects, and $\boldsymbol{\epsilon}_i$ is a within-individual residuals vector. The $\boldsymbol{\epsilon}_i$ are assumed to be indipendent and normally distributed with mean $\mathbf{0}$ and $n_i \times n_i$ variance-covariance matrix $\sigma_\epsilon^2 I_{n_i}$, where I_n denotes the $n \times n$ identity matrix. The random effects θ_i are assumed to be normally distributed with mean $\mathbf{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{\Sigma}$. The $\boldsymbol{\theta}_i$ are distributed indipendently of each other and of the within-subjects residuals $\boldsymbol{\epsilon}_i$. t_i is the survival time or some monotonic transformation of survival time such as the log of survival time, ξ is a $k \times 1$ vector of unknown parameters, w_i is a $k \times 1$ design matrix linking t_i to ξ, and finally λ is a $q \times 1$ vector of unknown parameters linking θ_i to t_i. r_i are assumed to be indipendent and normally distributed with mean 0 and variance-covariance matrix σ_r^2. The longitudinal marker and survival times are assumed independent conditional to random effects. In order to get the estimates for unknown parameters they developed an EM algorithm (Dempster *et al.*, 1977), a technique which iterates between solving for the expected values of functions of the unobserved data given the observed data and the maximum likelihood estimates of the parameters until convergence. Yet the dependence between longitudinal and survival process is not clear, since it is expressed by the random effects alone, and not by a parameter

linking the two processes directly.
Wulfsohn and Tsiatis (1997) model the marker's process by a liner growth curve model with random intercept and slope, given by

$$y_{ij} = \theta_{0i} + \theta_{1i} t_{ij} + \epsilon_{ij}$$

where ϵ_{ij} is normally distributed with mean 0 and variance σ^2_ϵ, $cov(\epsilon_{ij}, \epsilon_{ij'}) = 0$ for $j \neq j'$, the error is independent of the intercept and slope, and θ_{0i}, θ_{1i} are distributed as a bivariate normal with mean (θ_0, θ_1) and variance-covariance structure Σ. The survival model is given by a proportional hazards model

$$\lambda(t) = \lambda_0(t) exp\{\phi(\theta_{0i} + \theta_{1i} t)\}$$

where $\lambda_0(t)$ is the baseline hazard function at time t. Now the failure's hazard is function of "true" marker's value at each time t. Maximum likelihood estimates of all parameters are obtained by an implemented EM algorithm.
Henderson *et al.* (2000) propose to model the joint distribution of the measurements and the events for ith subject by a latent zero-mean bivariate Gaussian process $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$, which is realized indipendently in each individual. Hence, the joint model consists of two linked submodels:

1. measurement model for the longitudinal process, given by a random effects model:
$$y_{ij} = \mu_{1i}(t_{ij}) + W_{1i}(t_{ij}) + \epsilon_{ij}$$
where $\mu_{1i}(t) = \boldsymbol{x}_{1i}(t)\boldsymbol{\alpha}_1$ is the mean response, $\boldsymbol{x}_{1i}(t)$ and α_1 represent possibly time-dependent explanatory variables and their regression coefficients respectively, and $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ is the process of mutually independent measurement errors
2. intensity model for the survival process, given by a semiparametric proportional hazards model:
$$\lambda_i(t) = \lambda_0(t) exp\{\boldsymbol{x}_{2i}(t)\boldsymbol{\alpha}_2 + W_{2i}(t)\}$$
where $\lambda_0(t)$ is the baseline hazard function at time t, $\boldsymbol{x}_{2i}(t)$ and $\boldsymbol{\alpha}_2$ are possibly time-dependent explanatory variables and their regression coefficients respectively, including or not elements in common with $\boldsymbol{x}_{1i}(t)$.

Association between the longitudinal and survival process can arise in two ways: through common explanatory variables or through stochastic dependence between W_{1i} and W_{2i}. Hence, the longitudinal and survival process are conditionally independent, given W_1, W_2, and Z. They discuss special cases of this model class and extend the EM algorithm described by Wulfsohn and Tsiatis (1997). Yet, likewise

the model proposed by DeGruttola and Tu (1994), the dependence between the longitudinal and survival process is not expressed directly.

Song *et al.* (2002) consider the model of Wulfsohn and Tsiatis (1997), but relaxing the assumption of normality for the random effects, while requiring that θ_i have density belonging to a class of smooth densities studied by Gallant and Nichka (1987). The densities in this class are sufficiently differentiable to rule out behaviour such as jumps or oscillations and may be skewed, multimodal, and fat- or thin-tailed relative to the normal density, which is also belonging to this class (Zhang and Davidian, 2001).

Several Bayesian approaches have been developed too. Using those methods can be advantageous because, although computationally intensive, they make possible to fit the model without any asymptotic approximation, accomodate a variety of other expanded models, and their computational implementation is typically easier. Furthermore, with noninformative priors it is possible to mimic a corresponding likelihood analysis, where the likelihood is restandardized and interpreted as probability distribution on the parameters.

Faucett and Thomas (1996) propose the same random effects and proportional hazards models as Wulfsohn and Tsiatis (1997) by a Bayesian approach. In order to approximate likelihood methods, they use uninformative priors on all the parameters and they estimate the joint posterior distribution of all unknown parameters using Gibbs Sampling. Specifically they use flat priors for θ_0, θ_1, and γ, $|\Sigma|^{-3/2}$ for Σ, $1/\sigma_\epsilon^2$ for σ_ϵ^2, and $1/\lambda_{0j}$ for piecewise costant baseline hazard function λ_{0j}.

Berzuini and Larizza (1996) merge time series and failure time modeling within the theory of hierarchical models introduced by Lindley and Smith (1972). They consider time series data generated by a latent autoregressive stochastic process, allowing for smooth random fluctuations of each subject specific response around a linear trend. Then they extend the model to failure time data, by casting these data into the form of counts of failures for each subject in a sequence of time intervals. The relationships between the longitudinal and survival process are modeled by allowing the parameters that underlie each subject's time series to act as regressors in a Poisson regression model for the failure counts. They use Markov chain Monte Carlo methods for computing inferences in Bayesian analysis.

Wang and Taylor (2001) use a longitudinal model that incorporates a mean structure dependent on covariates, a random intercept, an integrated Ornstein-Uhlenbeck (IOU) stochastic process, and measurement error. By the parameters defining the IOU process it is possible to control the amount of smoothess of a person's path without imposing specific deterministic shapes on the path. A feature of this model is that the IOU process represents a family of covariance structures

with a random effects model and Brownian motion as special cases. The regression model for the event time data is a proportional hazards model that includes the longitudinal covariate as a time-dependent variable and other covariates. They used Bayesian techniques to fit the model.
Guo and Carlin (2004) develop the same model as Henderson *et al.* (2000) by a Bayesian approach, implemented via Markov chain Monte Carlo methods. They apply their method to a clinical trial and they compare the results to those obtained from readily available alternatives in SAS as well as Bayesian analogues of these traditional separate likelihood methods. The joint Bayesian approach appears to offer significantly improved and enhanced estimation of the parameters of interest, as well as simpler coding and comparable runtimes.

3.2 Extension to competing risks

The previous works have primarily focused on a single failure type with a non-informative censoring for the survival process. From a part, highlighting the survival process, we may be interested in modeling the time to occurrence of first failure, when several failure types are possible, from the other, we could deal with dependent censoring as competing event in order to model the longitudinal process correctly, when disease-related dropouts are evident. In both cases it is required the extension to a competing risks framework.
Elashoff *et al.* (2007) consider joint modelling of repeated measurements and competing risks failure time data to allow for more than one distinct failure type in the survival endpoint. They used a linear mixed effects model for longitudinal measurements and a mixture model for the survival process, similar to that of Larson and Dinse (1985b) and Ng and McLachlan (2003), but with the random effects. The mixture model for competing risks enables one to evaluate the effects of some factors on both the marginal probabilities of occurrence of the risks and the conditional cause-specific hazards, defined by a logistic and a proportional hazards model respectively, as follows:

$$P(L=l) = \frac{exp(\phi_{0l} + \boldsymbol{x}_i^{'}\boldsymbol{\phi}_{1l} + w_{li})}{1 + \sum_{l=1}^{k-1} exp(\phi_{0l} + \boldsymbol{x}_i^{'}\boldsymbol{\phi}_{1l} + w_{li})}, l = 1, \ldots, k-1$$

$$lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < T \leq t + \delta t | T > t, L = l) = \lambda_{0l} exp(\boldsymbol{x}_i^{''}(t)\boldsymbol{\gamma}_l + v_{li}), l = 1, \ldots, k$$

where $\boldsymbol{x}_i^{'}$ and $\boldsymbol{x}_i^{''}$ are two vectors of covariates, which may have factors in common, w_{il} and v_{il} are random effects for the lth risk, and λ_{0l} is an unspecified baseline

hazard function for risk l. The longitudinal measurements are independent of the competing risks, conditional on all the covariates and random effects. An EM-based algorithm is derived to obtain the parameter estimates. Since the estimation procedure is very complicated due to two-step mixture model for competing risks, additional hidden variables are needed to simplify the algorithm.
Hence we model jointly the longitudinal and survival process extended to a competing risks setting, by a linear mixed effects model and a semiparametric proportional hazards model respectively, first by a classical approach and then by a Bayesian one.

3.2.1 Notation

Let $\{y_{ij} : j = 1, \ldots, n_i\}$ the set of longitudinal quantitative measurements for the subject $i = 1 \ldots, m$. We suppose y_{ij} to be the biomarker of disease's progression measured at time t_j for the ith subject. The number and the time of measurements of the biomarker may be different for each individual. A linear mixed effects model is assumed for longitudinal response process

$$\begin{cases} \boldsymbol{y}_i = \boldsymbol{u}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{u}_i = \boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{z}_i\boldsymbol{\theta}_i \end{cases} \tag{3.1}$$

where $\boldsymbol{y}_i$ is a $n_i \times 1$ vector of repeated measurements on the ith subject, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are known design matrices, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown fixed effects, $\boldsymbol{\theta}_i$ is a $q \times 1$ vector of unobservable random effects, and $\boldsymbol{\epsilon}_i$ is a within-individual residuals vector. The $\boldsymbol{\epsilon}_i$ are assumed to be indipendent and normally distributed with mean $\mathbf{0}$ and $n_i \times n_i$ variance-covariance matrix $\sigma^2_\epsilon I_{n_i}$, where I_n denotes the $n \times n$ identity matrix. It implies that the n_i responses on subject i are independent, conditional on fixed effects $\boldsymbol{\alpha}_i$ and the random effects $\boldsymbol{\theta}_i$. The random effects θ_i are assumed to be normally distributed with mean $\mathbf{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{\Sigma}$. The $\boldsymbol{\theta}_i$ are distributed indipendently of each other and of the within-subjects residuals $\boldsymbol{\epsilon}_i$. Under this model $\boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{z}_i\boldsymbol{\theta}_i$ can be thought of as the true values of marker over time and the correlation between the repeated meaurements on an individual arises from an individual's deviation from overall effect. Marginally, the vector $\boldsymbol{y}_i$ is normally distributed with mean $\boldsymbol{x}_i\boldsymbol{\alpha}$ and variance-covariance matrix $\Sigma_\epsilon + \boldsymbol{z}_i\Sigma\boldsymbol{z}_i'$. The competing event time data for subject i is denoted by (t_i, δ_i, l_i), where $\delta_i = 0$ if the subject is censored, $\delta_i = 1$ if the subject is failed, t_i is the censoring time ($\delta_i = 0$) or the failure time ($\delta_i = 1$), and l_i is the failure type. A proportional hazards model is assumed for the lth event time

$$\lambda_l(t) = \lambda_{0l}(t) exp\Big\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_l + u_i(t)\gamma_l\Big\} \tag{3.2}$$

where $\lambda_{0l}(t)$ represents the baseline hazard, $u_i(t)$ is the true value of the biomarker, and $\boldsymbol{w}_i(t)$ denotes the vector of further covariates, which could include some or all of the $\boldsymbol{x}_i$ covariates. We assume that censoring, covariate errors, and timing of measurements are noninformative.
The longitudinal process and the survival process for lth failure type are linked by the parameter γ_l. In absence of association between the two processes the joint analysis should recover the same results as would be obtained from separate analysis for each component, and then γ_l should be equals to 0. Furthermore, since the joint model allows to evaluate the effect of each factor on both longitudinal and survival processes simultaneously, it is feasible to assess whether the effect of a factor on the competing event time is due only to its effect on the biomarker.

3.2.2 EM-based algorithm

In order to evaluate the effect of biomarker on survival outcome, it is needed to define the joint likelihood as the product of the likelihood of the longitudinal process multiplied by the likelihood of time to competing events conditional on the longitudinal process. The observed data for each individual is $(\boldsymbol{y_i}, \boldsymbol{x}_i(t), \boldsymbol{z}_i(t), t_i, \delta_i, \boldsymbol{w}_i(t))$ and the vector containing the unknown parameters is $\Omega = \{\boldsymbol{\alpha}, \Sigma, \Sigma_\epsilon, \lambda_{0l}(t), \gamma_l, \boldsymbol{\phi}_l\}$, $l = 1, \ldots, k$. The random effects $\boldsymbol{\theta}_i$ are not observable. The joint likelihood function for $\boldsymbol{\Omega}$, conditional on the observed data is given by:

$$L(\boldsymbol{\Omega}|t, \delta, \boldsymbol{y}) \propto \prod_{i=1}^{m} f(t_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{\Omega}) = \prod_{i=1}^{m} \int_{-\infty}^{\infty} f(t_i, \delta_i|\boldsymbol{y}_i, \boldsymbol{\Omega}, \boldsymbol{\theta}_i) f(\boldsymbol{y}_i|\boldsymbol{\Omega}, \boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i|\boldsymbol{\Omega}) d\boldsymbol{\theta}_i$$

$$= \prod_{i=1}^{m} \Big[\int_{-\infty}^{+\infty} \Big\{ \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\alpha}, \boldsymbol{\theta}_i, \sigma_\epsilon^2) \Big\} f(t_i, \delta_i|\lambda_{0l}, \boldsymbol{\phi}_l, \gamma_l, \boldsymbol{\alpha}, \boldsymbol{\theta}_i,) f(\boldsymbol{\theta}_i|\Sigma) d\boldsymbol{\theta}_i \Big]$$

$$= \prod_{i=1}^{m} \Big[\int_{-\infty}^{\infty} \Big\{ \lambda_{0l_i}(t_i) exp\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_{l_i} + u_i(t)\gamma_{l_i} \Big\}^{\delta_i} \times$$

$$\times exp\Big\{ - \int_0^{t_i} \sum_{l=1}^{k} [\lambda_{0l}(u) exp\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_l + u_i(t)\gamma_l\}] \Big\} \times$$

$$\times \frac{1}{\sigma_\epsilon^{n_i}} exp\Big[- \frac{1}{\sigma_\epsilon^2} \sum_{j=1}^{n_i} (y_{ij} - u_{ij})^2 \Big] \times |\Sigma|^{-1/2} exp\Big(- \frac{1}{2} \boldsymbol{\theta_i}' \Sigma^{-1} \boldsymbol{\theta_i} \Big) d\boldsymbol{\theta_i} \Big] \tag{3.3}$$

Since maximizing is difficult in the presence of integration, we can use the EM-based algorithm, which involves iterations between an E-step and an M-step. In order to maximize the conditional likelihood, the functions of $\boldsymbol{\theta_i}$ are replaced by their expectations given the observed data until the convergence. For simplicity we

suppose that the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}_l$ have only one component, hence the development to more components is obvious. The conditional expectation of the complete log-likelihood can be splitted in three pieces.
The conditional expectation of survival component is:

$$E_i\Big\{log\Big[\prod_{i=1}^{m}\lambda_{l_i}(t_i)^{\delta_i}exp\big(-\int_0^{t_i}\sum_{l=1}^{k}\lambda_l(u)du\big)\Big]\Big\}$$

By using the Cox model's properties, upon rearrangement of likelihood, the factor for failure l is:

$$E_i\Big\{log\Big[\prod_{i:L_i=l}\lambda_l(t_i)exp\big(-\sum_{i=1}^{m}\int_0^{t_i}\lambda_l(v)dv\big)\Big]\Big\} = E_i\Big\{\sum_{i:L_i=l}log(\lambda_l(t_i))+$$

$$-\sum_{i=1}^{m}\int_0^{t_i}\lambda_l(v)dv\Big\} = E_i\Big\{\sum_{i:L_i=l}\Big[log\big(\lambda_{0l}(t_i)\big)+w_i(t)\phi_l+u_i(t)\gamma_l\Big]+$$

$$-\sum_{i=1}^{m}\Big[\int_0^{t_i}\lambda_{0l}(v)exp\{w_i(t)\phi_l+u_i(t)\gamma_l\}dv\Big]\Big\}$$

$$=\sum_{i:L_i=l}log\big(\lambda_{0l}(t_i)\big)+\sum_{i:L_i=l}\Big(w_i(t)\phi_l+\gamma_l\big(\boldsymbol{x}_i\boldsymbol{\alpha}+E_i\big[\boldsymbol{z}_i\boldsymbol{\theta}_i\big]\big)\Big)$$

$$-\sum_{i=1}^{m}\int_0^{t_i}\lambda_{0l}(v)E_i\Big[exp\{w_i(t)\phi_l+u_i(t)\gamma_l\}\Big]dv\Big\}$$

Differentiating with respect to $\lambda_{0l}(v)$, we get:

$$\sum_{i=1}^{m}\Big\{\frac{I(L_i=l,t_i=v)}{\lambda_{0l}(v)}-E_i\Big[exp\{w_i(t)\phi_l+u_i(t)\gamma_l\}\Big]I(t_i>=v)\Big\}$$

Hence:

$$\widehat{\lambda}_{0l}(v)=\sum_{i=1}^{m}\frac{I(L_i=l,t_i=v)}{\sum_{j\in R(v)}E_j\Big[exp\{w_j(t)\phi_l+u_j(t)\gamma_l\}\Big]}$$

where the baseline hazard is calculated at each of the failure times and $R(v)$ is the set of subjects at risk at time v.
Differentiating with respect to ϕ_l gives:

$$\sum_{i=1}^{m}\Big\{I(L_i=l)w_i(t)-\int_0^{t_i}\lambda_{0l}(v)E_i\big[w_i(t)exp\{w_i(t)\phi_l+u_i(t)\gamma_l\}\big]I(t_i>=v)dv\Big\},$$

differentiating with respect to γ_l gives:

$$\sum_{i=1}^{m}\Big\{I(L_i=l)(E_i[u_i(t)])+$$

$$-\int_0^{t_i} \lambda_{0l}(v) E_i\big[u_i(t) exp\{w_i(t)\phi_l + u_i(t)\gamma_l\}\big] I(t_i >= v) dv\Big\}$$

Since $\lambda_{0l}(v)$ is a function of coefficients ϕ_l, and γ_l, there is not a closed-form solution to those equations. These parameters can be calculated conditional on the most updated values of other parameters, by using the Newton-Raphson algorithm in each iteration (Tjalling, 1995).
The conditional expectation of longitudinal component is:

$$E\Big[log \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij})\Big] = E \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Big[-\frac{1}{2} log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}(y_{ij} - x_i\alpha + \boldsymbol{z}_i\boldsymbol{\theta}_i)^2\Big] =$$

$$= -\frac{1}{2} \sum_{i=1}^{m} n_i log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E_i(y_{ij} - x_i\alpha + \boldsymbol{z}_i\boldsymbol{\theta}_i)^2$$

Differentiating with respect to σ_ϵ^2 gives:

$$\frac{d}{d\sigma_\epsilon^2} = -\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{m} n_i + \frac{1}{2\sigma_\epsilon^4} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E_i(y_{ij} - x_i\alpha + \boldsymbol{z}_i\boldsymbol{\theta}_i)^2$$

then

$$\widehat{\sigma}_\epsilon^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{E_i(y_{ij} - x_i\alpha + \boldsymbol{z}_i\boldsymbol{\theta}_i)^2}{\sum_{i=1}^{m} n_i}$$

Differentiating with respect to α gives:

$$\frac{d}{d\alpha} = \frac{1}{\sigma_\epsilon^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_i[y_{ij} - x_i\alpha - E_i(\boldsymbol{z}_i\boldsymbol{\theta}_i)]$$

then

$$\widehat{\alpha} = \sum_{i=1}^{m} \sum_{1}^{n_i} \frac{x_i[y_{ij} - E_i(\boldsymbol{z}_i\boldsymbol{\theta}_i)]}{\sum_{i=1}^{m} n_i x_i^2}$$

The conditional expectation of stochastic process is:

$$E\Big\{log \prod_{i=1}^{m} f(\theta|\Omega)\Big\} = E\Big\{log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi|\Sigma|}} exp\Big(-\frac{1}{2}\boldsymbol{\theta}' \Sigma^{-1}\boldsymbol{\theta}\Big)\Big\} =$$

$$= -\frac{m}{2} log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{m} E_i[(\boldsymbol{\theta}_i)' \Sigma^{-1}(\boldsymbol{\theta}_i)]$$

Differentiating with respect to Σ gives:

$$\widehat{\Sigma} = \sum_{i=1}^{m} \frac{E_i(\boldsymbol{\theta}_i\boldsymbol{\theta}_i')}{m}$$

The maximum likelihood estimate for $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \sum_{i=1}^{m} E_i(\boldsymbol{\theta_i})/m$$

In the expectation step of the $(m+1)$th iteration, we calculate $E[h(\boldsymbol{\theta}_i)|t_i, \delta_i, x_i, \boldsymbol{z}_i, w_i, \widehat{\Gamma}]$ where $\widehat{\Gamma}$ denotes the set of parameters estimated in the maximization step, that is $\widehat{\Gamma} = \{\alpha, \boldsymbol{\theta}, \Sigma, \sigma_\epsilon^2, \lambda_{0l}, \gamma_l, \phi_l\}, l = 1 \ldots, k$.
The conditional density of θ_i, given the observed data and the current parameters' estimate is equal to

$$f(\theta_i|t_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i, \widehat{\Gamma}) = \frac{f(\theta_i, t_i, \delta_i, l|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i, \widehat{\Gamma})}{f(t_i, \delta_i, l|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i, \widehat{\Gamma})} =$$

$$= \frac{f(t_i, \delta_i, l|\theta_i, \widehat{\lambda}_{0l}, \widehat{\phi}_l, \widehat{\gamma}_l) f(\theta_i|x_i, z_i, \widehat{\alpha}, \widehat{\theta}, \widehat{\Sigma}, \widehat{\sigma}_\epsilon^2)}{\int_{-\infty}^{\infty} f(t_i, \delta_i, l|\theta_i, \widehat{\lambda}_{0l}, \widehat{\phi}_l, \widehat{\gamma}_l) f(\theta_i|x_i, z_i, \widehat{\alpha}, \widehat{\theta}, \widehat{\Sigma}, \widehat{\sigma}_\epsilon^2) d\theta_i}$$

Hence $E[h(\boldsymbol{\theta}_i)|t_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{w}_i, \widehat{\Gamma}]$ is given by

$$\frac{\int_{-\infty}^{\infty} h(\theta_i) f(t_i, \delta_i, l|\theta_i, \widehat{\lambda}_{0l}, \widehat{\phi}_l, \widehat{\gamma}_l) f(\theta_i|x_i, z_i, \widehat{\alpha}, \widehat{\theta}, \widehat{\Sigma}, \widehat{\sigma}_\epsilon^2) d\theta_i}{\int_{-\infty}^{\infty} f(t_i, \delta_i, l|\theta_i, \widehat{\lambda}_{0l}, \widehat{\phi}_l, \widehat{\gamma}_l) f(\theta_i|x_i, z_i, \widehat{\alpha}, \widehat{\theta}, \widehat{\Sigma}, \widehat{\sigma}_\epsilon^2) d\theta_i}$$

The density $f(t_i, \delta_i, l|\theta_i, \widehat{\lambda}_{0l}, \widehat{\phi}_l, \widehat{\gamma}_l)$ has been defined in model 3.3 and the density $f(\theta_i|x_i, z_i, \widehat{\alpha}, \widehat{\theta}, \widehat{\Sigma}, \widehat{\sigma}_\epsilon^2)$ is a multivariate normal. The expectation of any function of θ_i can be calculated by using numerical integration (Press *et al.*, 1992).
By this procedure it is clear the difference between the two-stages and the joint model, because we use the information given by survival and longitudinal process simultaneously, by estimating the parameters defining the survival process as function of those defining the longitudinal process and viceversa.

3.2.3 Bayesian approach

One of the most used models for semiparametric survival analysis is the piecewise constant hazard model, that is

$$\lambda_{0l}(t) = \lambda_{jl} \qquad t_{j-1} \leq t < t_j, \qquad j = 1, \ldots, J$$

where t_{j-1}, t_j, $j = 0, \ldots, J$ define the intervals for $\lambda_{0l}(t)$. The more intervals there are, the more smoothing the baseline hazard function is, but the more parameters to estimate there are. The choice of the endpoints of those intervals is not based on the time of biomarker meaurements nor on the event time, it could be based on the quantiles of the observed time-to-events, in a such way that the intervals are

equally spaced or have an equal number of events. This semiparametric model is also known as piecewise exponential model, is quite general and can accomodate variuos shapes of the baseline hazard over the intervals. Now we can rewrite subject i's contribution to the joint likelihood function as

$$f(y_i, t_i, \delta_i, l_i) = \lambda_{0l_i}(t_i)^{\delta_i} exp\Big\{\delta_i\big(\boldsymbol{w}_i(t)\boldsymbol{\phi}_{l_i} + u_i(t)\gamma_{l_i}\big) +$$

$$-\sum_{l=1}^{k}\sum_{j=1}^{J}(I(t_i > t_{j-1})\lambda_{0lj}\int_{t_{j-1}}^{min(t_j,t_i)} exp\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_l + u_i(t)\gamma_l\}du\Big\}\times$$

$$\times\frac{1}{(2\pi\sigma_\epsilon^2)^{m_i/2}}exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{j=1}^{n_i}\{y_{ij} - u_{ij}\}^2\Big)$$

In order to make possible a fair comparison of the classical and Bayesian analysis, we select vague prior distributions, that is we use proper priors, but with hyperparameter values chosen so that the priors will have a minimal impact on data. Specifically, in the longitudinal submodel we use multivariate normal priors for the fixed effects vector $\boldsymbol{\alpha}$ and an inverse gamma priors for the error variance σ_ϵ^2, both having very low precision. In the survival submodel we use the conjugate prior for underlying hazard, $\lambda_{jl} \sim Gamma(p_{0jl}, q_{0jl})$ for $j = 1, \ldots, J$, where $Gamma(p_{0jl}, q_{0jl})$ denotes the gamma distribution with shape parameter p_{0jl} and scale parameter q_{0jl}. Here p_{0jl}, q_{0jl} are prior parameters which can be elicited by the prior mean and variance of λ_{jl}. We take vague normal priors for $\boldsymbol{\phi}_l$ and γ_l. Finally for the parameters common to both models, Σ, we select a inverse Wishart, because it allows for a good identifiability of the main effects, providing some shrinkage of the random effects towards 0 (Carlin and Luis, 2000).

Let $[\cdot]$ and $[\cdot|\cdot]$ be the marginal and conditional density respectively. The likelihood multiplied by the priors for the longitudinal model is given by

$$\prod_{i=1}^{m}\Big(\Big[\prod_{j=1}^{n_i}\left(y_{ij}|\boldsymbol{\alpha}, \boldsymbol{\theta}_i, \sigma_\epsilon^2, \Sigma\right)\Big]\Big)\Big[\boldsymbol{\alpha}\Big]\Big[\boldsymbol{\theta}_i|\Sigma\Big]\Big[\sigma_\epsilon^2\Big]\Big[\Sigma\Big]$$

The likelihood multiplied by the priors for the survival model in presence of competing risks is given by

$$\prod_{i=1}^{m}\Big(\Big[t_i, \delta_i|\lambda_{0l}, \boldsymbol{\phi}_l, \gamma_l\Big]\Big)\Big[\lambda_{0l}\Big]\Big[\boldsymbol{\phi}_l\Big]\Big[\gamma_l\Big]$$

$l = 1, \ldots, k$. The joint posterior distribution of all parameters is proportional to the product of the likelihood defined above. In order to obtain the posterior marginal distributions for the parameters, Markov chain Monte Carlo (MCMC) methods are used. The procedure consists of iterating through the parameters, either in

blocks or singly, and drawing from the appropriate full conditional distributions of each parameter, given the current assigment of all other parameters and data. When sampling from the full conditional distribution is not feasible, we use the Metropolis Hastings Sampling (Hastings, 1970). By using Bayes' rule, the posterior distribution of the random effects θ_i at the $(m+1)$th iteration is proportional to

$$\begin{aligned}&[\boldsymbol{\theta}_i|\boldsymbol{\alpha}^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}, \lambda_{0l}^{(m)}(t), \boldsymbol{\phi}_l^{(m)}, \gamma_l^{(m)}, \{\boldsymbol{y}_{ij}\}, t_i, \delta_i]\\ \propto\ &[\{y_{ij}\}|\boldsymbol{\theta}_i, \boldsymbol{\alpha}^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}] \times [\boldsymbol{\theta}_i|\boldsymbol{\alpha}^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}]\times\\ &\times[t_i, \delta_i|\boldsymbol{\theta}_i, \boldsymbol{\alpha}^{(m)}, \lambda_{0l}^{(m)}(t), \phi_l^{(m)}, \gamma_l^{(m)}]\end{aligned}$$

$l = 1, \ldots, k$ and for $i = 1, \ldots, m$.
The first term is a normal distribution proportional to

$$N\Big(\boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{z}_i\boldsymbol{\theta}_i, \sigma_\epsilon^2 I_{n_i}\Big)$$

The second term is a normal distribution proportional to

$$N\Big((\boldsymbol{z}_i^{'}(\sigma_\epsilon^2 I_{n_i})^{-1}\boldsymbol{z}_i + \Sigma)^{-1}\boldsymbol{z}_i^{'}(\sigma_\epsilon^2 I_{n_i})^{-1}(y_i - \boldsymbol{x}_i\boldsymbol{\alpha}), (\boldsymbol{z}_i^{'}(\sigma_\epsilon^2 I_{n_i})^{-1}\boldsymbol{z}_i + \Sigma^{-1})^{-1}\Big)$$

The third term is the full likelihood of the survival parameters

$$\Big\{\lambda_{0l_i}(t_i)exp\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_{l_i} + u_i(t)\gamma_{l_i}\Big\}^{\delta_i} exp\Big\{-\int_0^{t_i}\sum_{l=1}^{k}\big[\lambda_{0l}(v)exp\{\boldsymbol{w}_i(v)\boldsymbol{\phi}_l + u_i(v)\gamma_l\}\big]dv\Big\}$$

To obtain the posterior distribution of the fixed effects at the $(m+1)$th Gibbs iteration, given by

$$\begin{aligned}&[\boldsymbol{\alpha}|\boldsymbol{\theta_i}^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}, \lambda_{0l}^{(m)}(t), \boldsymbol{\phi}_l^{(m)}, \gamma_l^{(m)}, \{\boldsymbol{y}_{ij}\}, t_i, \delta_i] \propto\\ &[\{y_{ij}\}|\boldsymbol{\theta}_i, \boldsymbol{\alpha}^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}] \times [\boldsymbol{\alpha}|\boldsymbol{\theta}_i^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}]\times\\ &\times[t_i, \delta_i|\boldsymbol{\theta}_i, \boldsymbol{\alpha}^{(m)}, \lambda_{0l}^{(m)}(t), \boldsymbol{\phi}_l^{(m)}, \gamma_l^{(m)}]\end{aligned}$$

$l = 1, \ldots, k$, it is needed to redefine only the second term of the posterior distribution of the random effects, being a normal distribution

$$N\Big(\big(\sum_{i=1}^{m}\boldsymbol{x}_i^{'}(\sigma_\epsilon^2 I_{n_i})^{-1}\boldsymbol{x}_i\big)^{-1}\big(\sum_{i=1}^{m}\boldsymbol{x}_i^{'}(\sigma_\epsilon^2 I_{n_i})^{-1}(y_i - \boldsymbol{z}_i\boldsymbol{\theta}_i)\big), \big(\sum_{i=1}^{m}\boldsymbol{x}_i^{'}(\sigma_\epsilon^2 I_{n_i})^{-1}\boldsymbol{x}_i\big)^{-1}\Big)$$

For parameters $\Sigma, \sigma_\epsilon^2, \lambda_{0l}, \phi_l, \gamma_l$, $l = 1, \ldots, k$, each of their conditional distributions is a product of a standard distribution obtained from the likelihood and the prior. Let $\Sigma \sim IW(s, S)$ with degrees of freedom $s \geq q$ and $q \times q$ covariance matrix, hence the conditional distribution of Σ is

$$[\Sigma|\cdot] \propto [\{y_{ij}\}|\boldsymbol{\theta}_i, \boldsymbol{\alpha}^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}] \times [\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \Sigma, \sigma_\epsilon^2] \times [\Sigma] \propto IW\Big(s+q, S+\sum_{i=1}^{n} \boldsymbol{\theta}_i \boldsymbol{\theta}_i^{'}\Big).$$

Let $\sigma_\epsilon^2 \sim IG(p, q)$, hence the conditional distribution of σ_ϵ^2 is

$$[\sigma_\epsilon^2|\cdot] \propto [\{y_{ij}\}|\boldsymbol{\theta}_i, \boldsymbol{\alpha}, \Sigma, \sigma_\epsilon^2] \times [\sigma_\epsilon^2] \propto$$

$$\propto IG\Big(\frac{\sum_{i=1}^{m} n_i}{2} - 1 + q, \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - (\boldsymbol{x}_i\boldsymbol{\alpha} - \boldsymbol{z}_i\boldsymbol{\theta}_i))^2}{2} + p\Big).$$

Let $\lambda_{0jl} \sim Gamma(p_{0jl}, q_{0jl})$, then the conditional distribution of λ_{0jl} is

$$[\lambda_{0jl}|\cdot] \propto [t_i, \delta_i|\boldsymbol{\theta}_i, \boldsymbol{\alpha}, \lambda_{0jl}, \boldsymbol{\phi}_l, \gamma_l] \times [\lambda_{0jl}] \propto$$

$$\propto G\Big(d_{l,k} + 1 + p_{0jl}, \sum_{i:t_i \geq t_j} \int_{t_{j-1}}^{t_j} exp\{\boldsymbol{w}_i(v)\boldsymbol{\phi}_l + u_i(v)\gamma_l\}dv +$$

$$\sum_{i:t_i \in (t_{j-1}, t_j), d_i = l} \int_{t_{j-1}}^{t_i} exp\{\boldsymbol{w}_i(v)\boldsymbol{\phi}_l + u_i(v)\gamma_l\}dv + q_{0jl}\Big)$$

The posterior distribution of parameters $\boldsymbol{\phi}_l$ and γ_l is respectively given by

$$[\boldsymbol{\phi}_l|\cdot] \propto [t_i, \delta_i|\boldsymbol{\theta}_i, \boldsymbol{\alpha}, \lambda_{0kl}, \boldsymbol{\phi}_l, \gamma_l] \times [\boldsymbol{\phi}_l],$$

$$[\gamma_l|\cdot] \propto [t_i, \delta_i|\boldsymbol{\theta}_i, \boldsymbol{\alpha}, \lambda_{0kl}, \boldsymbol{\phi}_l, \gamma_l] \times [\gamma_l]$$

Both distributions are proportional to

$$\Big\{\lambda_{0l_i}(t_i) exp\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_l + u_i(t)\gamma_l\Big\}^{\delta_i} exp\Big\{-\int_0^{t_i} \sum_{l=1}^{k} \big[\lambda_{0l}(u) exp\{\boldsymbol{w}_i(t)\boldsymbol{\phi}_l + u_i(t)\gamma_l\}\big]\Big\}$$

By assuming a normal prior for $\boldsymbol{\phi}_l$ and γ_l, we use Metropolis Hasting Sampling to draw a sample from this distribution.

Part III

Application

4

The CASCADE Study

4.1 Description of the data

The aim of this work is to develop an appropriate methodology to data arising from one of the largest AIDS multicentre studies, the CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe) Study, a collaboration representing 22 cohorts based in Europe, Australia and Canada. Unlike other studies in this field, the date of seroconversion of all participants is reliably estimated, the majority (81%) being the midpoint between the first positive and last negative antibody test dates with a maximum 3-year interval between test dates, the minority (19%) on the basis of laboratory evidence or seroconversion illnesses. In addition both their CD4 cell count and RNA viral load are recorded longitudinally from entry into the study till the end of their follow-up, and treatment history and AIDS-related events are carefully recorded. People aged under 15 years at seroconversion are excluded from all analyses as the definition of AIDS differs in children. Subjects with an AIDS diagnosis prior to entry, subjects who have received ART [1] before starting HAART or without at least two measurements of CD4 cell count and HIV RNA viral load during the study are excluded from the analysis. Since the progression of disease differs in individuals with different modality of infection, the study includes only homosexual men who are seroconverted since 1984 to 2005.
As introduced in chapter 1, the goal of this study is the evaluation of the risk to get AIDS in presence of competing events, that is, the interruption and the modification of therapy, for each group defined according to CD4 cell count at HAART initiation, lower than 200 cell/μl, included between 200 and 350 cell/μl, and higher than 350 cell/μl. In order to make it, first we need to estimate the CD4 cell count and viral load patterns over time since seroconversion, then to estimate the effect

[1] active antiretroviral therapy

of such patterns on time to competing events, adjusted for other covariates.
By the end of follow-up, 23 (2.11%) has developed AIDS, 279 (25.60%) has interrupted the treatment for at least one week, and 338 (31.01%) has changed the therapy. After identifying the possible causes of dependent censoring, such as the sospension and the modification of therapy, and considering them as competing events for AIDS, we assume an independent censoring meachanism for the withdrawals from the study for unknown reasons. We report the different times of seroconversion, HAART initiation, and exit from the study and the observed events for each group, in table 4.1 and 4.2 respectively.

Table 4.1. Characteristics according to CD4 at HAART initiation

Age at seroconversion	**median(IQR)**
<200	31 (25-36)
[200-350[	32 (27-39)
$\geq$ 350	32 (27-39)
Seroconversion year	**median(IQR)**
< 200	1996 (1993-2000)
[200-350[	1998 (1994-2001)
$\geq$ 350	1998 (1996-2001)
HAART year	**median(IQR):**
< 200	2001 (1998-2003)
[200-350[	2001 (1999-2004)
$\geq$ 350	2000 (1998-2002)
Elapsed years between SC and HAART	**median(IQR):**
< 200	4.33 (1.28-7.13)
[200-350[	2.78 (1.20-5.78)
$\geq$ 350	0.94 (0.19-2.76)

The larger the elapsed time between seroconversion and the initiation of therapy is, the lower the CD4 cell count is, and the higher the probability to get AIDS and to change therapy is. 5.97% of patients whose CD4 cell count is lower than 200 cell/μl at HAART initiation gets AIDS versus 1.96% and 1.20% of those whose CD4 cell count is included between 200 and 350 cell/μl and higher than 350 cell/μl, respectively. Contrary, 33.46% of patients whose CD4 cell count is higher than 350 cell/μl interrupts the treatment versus 19.07% and 13.43% of those whose CD4 cell count is included in [200,350) cell/μl and lower than 200 cell/μl, respectively. Finally 34.33% of patients whose CD4 cell count is lower than 200 cell/μl changes therapy versus 29.58% and 31.26% of those whose CD4 cell count is included between 200 and 350 cell/μl and higher than 350 cell/μl, respectively.
The data include a total of 15377 CD4 cell count and 9927 viral load measurements

Table 4.2. Failures of 1090 individuals

Failure	Number	(%)
<200	134	(12.29)
Censored	62	(46.27)
AIDS	8	(5.97)
Interruption of therapy	18	(13.43)
Change of therapy	46	(34.33)
[200-350[	409	(37.52)
Censored	202	(43.39)
AIDS	8	(1.96)
Interruption of HAART	78	(19.07)
Change of therapy	121	(29.58)
$\geq$ 350	547	(50.18)
Censored	186	(34.00)
AIDS	7	(1.28)
Interruption of HAART	183	(33.46)
Change of therapy	171	(31.26)

taken on patients. On average, 14 (range, 2-67) CD4 cell count measurements per subject are available with median interval of 92 days between any two successive CD4 cell count measurements. On average, 9 (range, 2-41) RNA viral load measurements per subject are available with median interval of 92 days between any two successive RNA viral load measurements. The CD4 cell count at seroconversion is known for 77% of individuals, while the viral load is unknown. CD4 cell count and viral load at the initiation of HAART, when unknown, are estimated by the mean of biomarkers' measurements recorded in the last six months before submission to therapy. 12.29% of patients starts HAART when CD4 cell count is lower than 200 cell/μl, 37.52% when CD4 cell count is included in [200, 350) cell/μl, finally 50.18% when CD4 cell count is higher than 350 cell/μl, as reported in table 4.3.

We transform the CD4 cell count and RNA viral load to a square-root and log_{10} scale respectively, in order to normalize the data. We show the plots of CD4 cell count and viral load measurements in figure 4.1 and 4.2. Obviously, the measurements number of both biomarkers decreases over the time, because of loss to follow-up for the occurrence of some competing risks or "natural" censoring, and reasonably the viral load values are more variable than CD4 cell count values.

4.2 Longitudinal and survival models

Since the clinical question our method wants to dress is at which stage it is better to start the therapy, that is, when CD4 cell count is included between 200 and 350 cell/μl rather than when it is higher than 350 cell/μl, or viceversa, by considering

Table 4.3. CD4 cell count and viral load

CD4 at seroconversion	
<200	37 (3.39)
[200-350[	152 (13.94)
≥ 350	645 (59.17)
Unknown	256 (23.49)
CD4 at HAART	
<200	134 (12.29)
[200-350[	409 (37.52)
≥ 350	547 (50.18)
RNA viral load at HAART	
<10000	606 (55.60)
≥ 10000	484 (44.90)

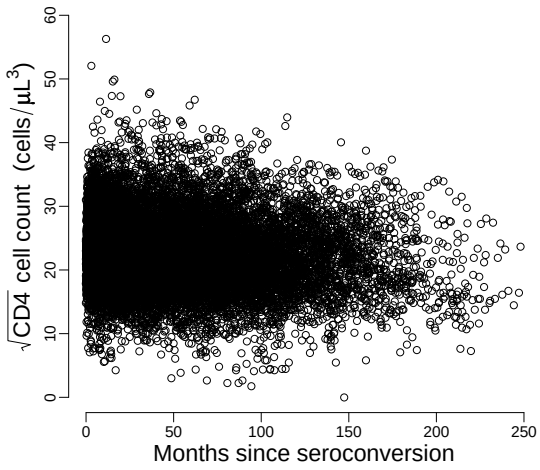

Fig. 4.1. Longitudinal measurements of CD4 cell count on the time since seroconversion

also the elapsed time between seroconversion and HAART initiation, we will fit the CD4 cell count pattern over time since seroconversion and we will compare the risks of failure for each group. We will proceed by steps: first we will model the CD4 cell count and the competing events separately, then by fitting the longitudinal data by an appropriate model we will include the fitted values in the survival model, and finally we will model the longitudinal and survival processes jointly. The last step will consist of inclusion of viral load in the analysis. We will distinguish two groups of patients, those whose viral load is lower than 10000 copies/mL and those whose

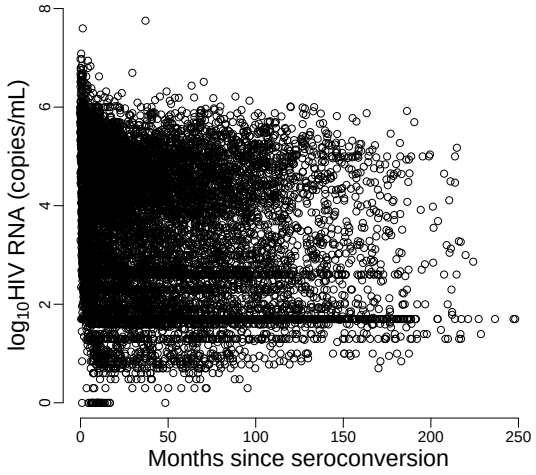

Fig. 4.2. Longitudinal measurements of HIV-RNA on the time since seroconversion

viral load is higher than 10000 copies/mL at HAART initiation. For an exploratory analysis, similarly to CD4 cell count, first we will model it separately, then jointly to CD4 cell count. Finally we will evaluate the effect of CD4 cell count and viral load on competing events by first including the fitted biomarkers' values in the survival model and then by modelling the processes jointly. However, because of computational complexity, due to large dimension of dataset, to unbalanced data, and to presence of multiple time-scales, before modelling the data jointly by a Bayesian approach, we will make a first analysis in order to select an appropriate class of models for the longitudinal and the survival data respectively, by traditional methods, i.e. by comparing AIC (Akaike Information Criterion) (Akaike, 1974) of several longitudinal models and by testing the proportionality assumption for survival data by Schoenfeld residuals.

4.2.1 Univariate longitudinal model and competing risks

The presence of multiple time-scales does not allow to represent significantly the biomarker pattern in a single plot because we observe at the same time patients without any therapy but also those just submitted to HAART but with different treatment histories. Hence it is needed to stratify by elapsed time between seroconversion and HAART initiation in order to visualize the presence of some trends. By

using the deciles of distribution of variable "elapsed time between seroconversion and HAART initiation" as stratification criteria, we represent the box-plots of CD4 cell count for each interval of time equal to two months and a smoothing function, which uses locally-weighted polynomial regression, remarking the interval of time the individuals start the treatment. Specifically, we show CD4 cell count pattern of individuals who start treatment between 94 and 122 months after seroconversion in figure 4.3. Similar plots are obtained for each strata. Approximately the biomarker

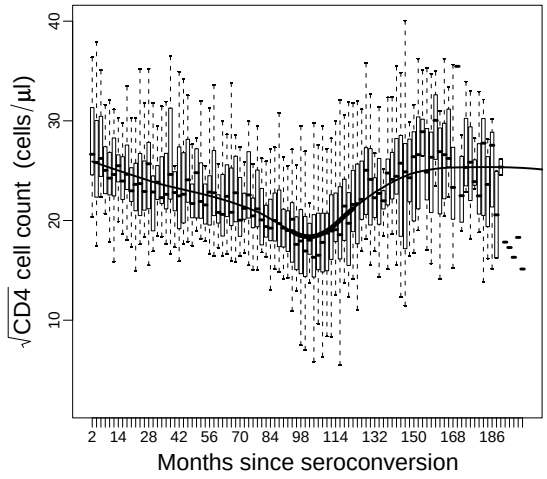

Fig. 4.3. CD4 cell count pattern of individuals, who start therapy between 94 and 122 months after seroconversion

decreases after seroconversion, reaching its minimum value before the HAART initiation, it increases rapidly in the first months after submission to therapy, and then it tends to stabilize.

We proceed by making a further stratification based on groups defined by different CD4 cell count at the HAART initiation in order to make a comparison between them. After choosing arbitrarily an interval of time of HAART initiation, we represent the CD4 cell count pattern by a smoothing function, for the individuals whose CD4 cell count is lower than 200 cell/μl, those whose CD4 cell count is included between 200 and 350 cell/μl, finally those whose CD4 cell count is higher than 350 cell/μl at HAART initiation, in figure 4.4.

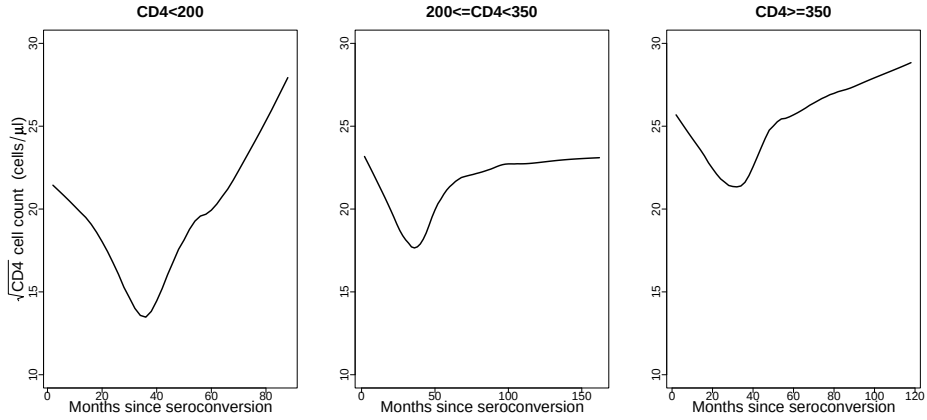

Fig. 4.4. CD4 cell count pattern represented for the groups of patients, who starts therapy in an interval of time included between 32 and 44 months after seroconversion and characterized by different CD4 cell count at HAART initiation, <200, [200, 350), ≥350 respectively.

At the first stage the therapy seems to increase the CD4 cell count almost as its level at seroconversion, then its effect attenuates after first months, mainly for the patients whose CD4 cell count is included between 200 and 350 cell/μl at the initiation of HAART. Obviously this result also depends on interval of time chosen to represent the different trends.

We use a piecewise linear mixed effects model with a first slope representing the CD4 cell count pattern before the submission to therapy, a second slope representing the short term response of CD4 cell count to therapy and a third slope representing the long term response. Being $\{y_{ij}^{'} : j = 1, \ldots, n_i^{'}\}$ the CD4 cell count measurements at times $\{t_{ij} : j = 1, \ldots, n_i^{'}\}$, the model is given by:

$$y_{ij}^{'} = u_{ij}^{'} + \epsilon_{ij}^{'} \tag{4.1}$$

$$\begin{aligned} u'_{ij} = {} & \alpha_1 + \alpha_2 I_1 + \alpha_3 I_2 + (\alpha_4 + \alpha_5 I_1 + \alpha_6 I_2) t_{ij} + (\alpha_7 + \alpha_8 I_1 + \alpha_9 I_2)(t_{ij} - t_{i1}) * \\ & I(t_{ij} - t_{i1} > 0) + (\alpha_{10} + \alpha_{11} I_1 + \alpha_{12} I_2)(t_{ij} - t_{i2}) I(t_{ij} - t_{i2} > 0) \\ & + \theta_{1i} + \theta_{2i} t_{ij} + \theta_{3i}(t_{ij} - t_{i1}) I(t_{ij} - t_{i1} > 0) + \theta_{4i}(t_{ij} - t_{i2}) I(t_{ij} - t_{i2} > 0) \end{aligned}$$

where $I_1 = 1$ if CD4 is included in [200, 350) cell/μl at therapy initiation otherwise 0, $I_2 = 1$ if CD4 is higher than 350 cell/μl otherwise 0, t_{i1} is the time of start of therapy dependent on ith subject, and t_{i2} is the time when the slope changes because of therapeutic effect's decrease, dependent on ith subject. We estimate this

time by adapting the model 4.1 with different values of t_{i2} and comparing the AIC values obtained. Results suggest to fix t_{i2} to three months after therapy initiation. $\boldsymbol{\alpha}$ is a 12×1 vector of unknown fixed effects, $\boldsymbol{\theta}_i$ is a 4×1 vector of unobservable random effects, and $\boldsymbol{\epsilon}_i$ is a within-individual residuals vector. The $\boldsymbol{\epsilon_i}$ are assumed to be normally distributed with mean $\mathbf{0}$ and $n_i^{'} \times n_i^{'}$ variance-covariance matrix $\sigma_\epsilon^2 I_{n_i'}$, where $I_{n'}$ denotes the $n^{'} \times n^{'}$ identity matrix. The random effects θ_i are assumed to be normally distributed with mean $\mathbf{0}$ and 4×4 unstructured variance-covariance matrix $\boldsymbol{\Sigma}$. The $\boldsymbol{\theta}_i$ are distributed indipendently of each other and of the within-subjects residuals ϵ_{ij}. Besides the model 4.1, specific submodels have been considered, in particular a quadratic term has been included, random effects have been excluded, several covariance structures have been applied, and splines have been adapted. In terms of a large likelihood, a small number of parameters, and a clear interpretation of the model, the best model is the model 4.1. By selecting three subjects randomly, we show the individual deviation from overall effect, due to the inclusion of random effects, in figure 4.5.

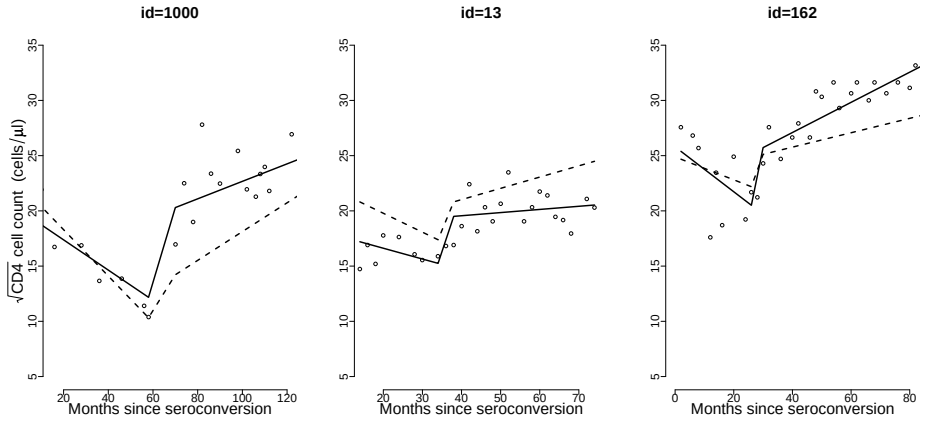

Fig. 4.5. Predicted individual (solid line) and mean (broken line) CD4 cell count pattern over time since seroconversion for three individuals selected randomly, whose CD4 cell count is lower than 200 cell/μl, is included in [200, 350) cell/μl, and is higher than 350 cell/μl at HAART inititiation, respectively.

We report the coefficients estimates, obtained by restricted maximum likelihood and Bayesian approach, in table 4.4. To allow for a fair comparison between classical and Bayesian analyses we use proper prior distributions but with hyperparameters values chosen so that the priors have a minimal impact on data. Specifically, we

take multivariate normal for the main effects vector $\boldsymbol{\alpha}$, an inverse gamma prior for the error variance σ_ϵ^2, and an inverse Wishart for Σ, all having very low precision. We monitor the MCMC convergence by three parallel MCMC sampling chains of 150000 iterations each, following a 50000-iteration "burn-in" period.

In the classical approach the RMLE reachs convergence only introducing in the

Table 4.4. Coefficients' estimates (CI(95%)) in piecewise linear mixed effects model

	Classical		**Bayesian**	
α_1	22.169	(21.279, 23.058)	22.280	(21.360, 23.200)
α_2	0.596	(-0.427, 1.619)	0.756	(-0.340, 1.820)
α_3	2.573	(1.579, 3.567)	2.502	(1.490, 3.532)
α_4	-0.309	(-0.346, -0.272)	-0.419	(-0.510, -0.327)
α_5	0.069	(0.026, 0.112)	0.077	(-0.032, 0.191)
α_6	0.175	(0.130, 0.220)	0.212	(0.101, 0.326)
α_7	1.364	(1.137, 1.591)	1.842	(1.491, 2.191)
α_8	0.404	(0.147, 0.661)	0.215	(-0.186, 0.623)
α_9	0.119	(-0.142, 0.380)	-0.171	(-0.567, 0.229)
α_{10}	-0.797	(-1.032, -0.562)	-1.162	(-1.540, -0.787)
α_{11}	-0.529	(-0.794, -0.264)	-0.349	(-0.770, 0.078)
α_{12}	-0.438	(-0.706, -0.170)	-0.171	(-0.606, 0.251)

model $\theta_{1i}, \theta_{2i}, \theta_{3i}$ but not θ_{4i}. The difference between results obtained by classical and Bayesian approach may be due to random effects θ_{4i}, which takes off significativity to α_{11}, α_{12}. On equal terms of elapsed time between seroconversion and start of treament, the CD4 cell count pattern does not differ significantly before starting therapy in individuals whose CD4 cell count is included between 200 and 350 cell/μl and those whose CD4 cell count is lower than 200 cell/μl at HAART initiation. At seroconversion both groups have lower CD4 cell count than that of individuals whose CD4 cell count is higher than 350 cell/μl at therapy initiation, and the CD4 cell count decrease in an unit of time is almost double in the first two groups compared to the third one. The therapeutic effect is clearly positive, also if after three months it attenuates, but it is not significantly different in the three groups. Consistently with the results, the covariance matrix of random effects is given by

$$\begin{pmatrix} 23.370 & -0.431 & -1.855 & 2.309 \\ -0.431 & 0.219 & -0.117 & 0.016 \\ -1.855 & -0.117 & 2.216 & -2.051 \\ 2.309 & 0.016 & -2.051 & 2.363 \end{pmatrix}$$

We represent the box-plots of the last CD4 cell count measurement for competing events, AIDS, interruption of therapy, and change of therapy respectively in figure 4.6. Obviously the individuals who get AIDS have a lower CD4 cell count than those

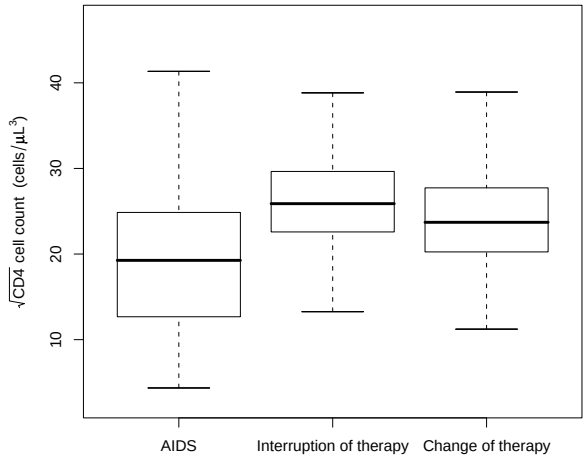

Fig. 4.6. Box-plot of last CD4 cell count measurement of individuals who get AIDS, interrupt the therapy, and change the therapy, respectively.

who "fail" from competing events. It is interesting to observe that the patients who interrupt the therapy have a lightly higher CD4 cell count than those who change the therapy. It is reasonable because the treatment interruption could be used as a strategy for boosting immune response to HIV or reducing long-term toxicity when the patient's immunological status is high enough, while a change to second-line therapy gets necessary if the first-line treatment fails, and then the CD4 cell count has not reached a satisfactory level. Yet, a complete analysis over the time since seroconversion is required to evaluate the effect of the biomarker pattern on the probability of failing from a competing event.

The non-parametric cumulative incidence curves are shown in figure 4.7.

The lower the CD4 cell count is at HAART initiation, the higher the probability to get AIDS is, and the lower the probability to interrupt the therapy is. The probability to change the therapy does not seem to be different between the groups. Since grafically the proportionality hypothesis is satisfied for each "failure", we

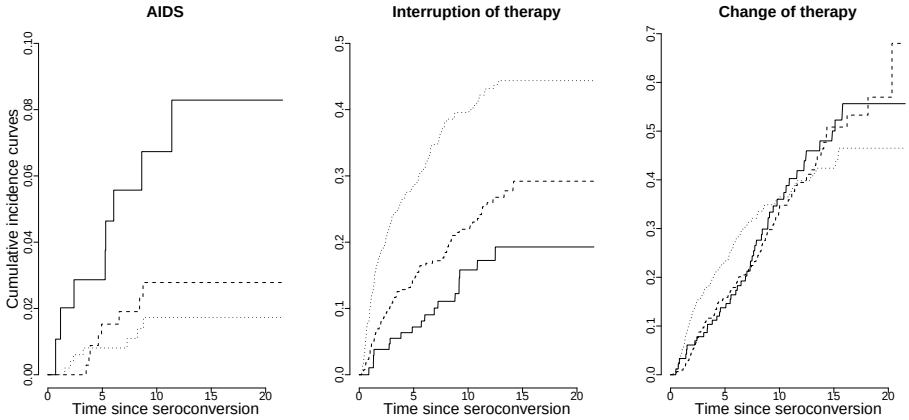

Fig. 4.7. Comparison between non-parametric cumulative incidence curves for individuals, whose CD4 cell count is lower than 200 cell/μl (solid line), those whose CD4 cell count is included in [200, 350) cell/μl (broken line), and those whose CD4 cell count is higher than 350 cell/μl (dotted line) at the therapy initiation.

model the hazard to fail from three different competing events by a proportional hazards model, given by

$$\lambda_l(t) = \lambda_{0l}(t) exp(\phi_{1l} w_1 + \phi_{2l} w_2 + \phi_{3l} w_3 + \phi_{4l} w_4) \tag{4.2}$$

where $l = 1, 2, 3$ indicates the competing event, AIDS, interruption of therapy, and change of therapy respectively, and $w_1 = 1$ if CD4 cell count at HAART initiation is included between 200 and 350 cell/μl, otherwise 0, $w_2 = 1$ if it is higher than 350 cell/μl, w_3 is the age at seroconversion, considered as continuos variable, and $w_4 = 1$ if the seroconversion year is after 1995, otherwise 0. The effect of the CD4 cell count at HAART initiation on the hazard functions is adjusted for age at seroconversion and seroconversion year. Since our dataset includes only men, who has got HIV by homosexual relationships, no more not hidden-variables can be possible confounders, other than viral load. We report the results in the table 4.5.

These results are in agreement with those obtained previously in a non-parametric setting. The CD4 cell count at HAART initiation significantly affects the hazard to fail from AIDS for both groups, on the hazard to interrupt the therapy only for the subjects whose CD4 cell count is higher than 350 cell/μl, while it does not affect significantly the hazard to change therapy. We check the proportionality assumption by Schoenfeld residuals and, besides model 4.2, we consider and compare by Wald tests models including iteractions between the variables w_1, w_2 and w_3.

Now, we extend the analysis, treating CD4 cell count as continuos variable, and

Table 4.5. Coefficients estimates by Cox proportional hazards model

	Estimate	95% Conf. Interval	p-value
ϕ_{11}	-1.17	(-2.15,-0.19)	0.019
ϕ_{21}	-1.60	(-2.55,-0.65)	0.001
ϕ_{31}	0.00	(-0.05,0.05)	0.981
ϕ_{41}	0.20	(-0.58,0.98)	0.622
ϕ_{12}	0.31	(-0.20,0.82)	0.235
ϕ_{22}	0.80	(-0.32,1.29)	0.001
ϕ_{32}	0.00	(-0.02,0.00)	0.329
ϕ_{42}	0.60	(0.30,0.89)	0.000
ϕ_{13}	-0.17	(-0.50,0.17)	0.314
ϕ_{23}	-0.07	(-0.40, 0.25)	0.663
ϕ_{33}	0.00	(-0.01,0.01)	0.914
ϕ_{43}	-0.01	(-0.24,0.21)	0.901

evaluating the effect of its pattern on the three competing events. We will compare two approaches: the two-stages model and the joint model. The first one consists of modeling first the longitudinal data, and then including the fitted values in an appropriate survival model. The second one estimates the parameters of longitudinal and survival models jointly. The longitudinal model is given by the piecewise linear mixed effects model, defined in 4.1. The survival process is modelled by a Cox proportional hazards model, given by

$$\lambda_l(t) = \lambda_{0l}(t) exp\{(\gamma_l + \gamma_{1l} I_1 + \gamma_{2l} I_2) u^{'}(t) + \phi_{1l} w_1 + \phi_{2l} w_2\} \tag{4.3}$$

where $I_1 = 1$ if CD4 cell count at HAART initiation is included between 200 and 350 cell/μl, otherwise 0, $I_2 = 1$ if CD4 cell count is higher than 350 cell/μl, otherwise 0, $u^{'}(t)$ is the biomarker value at time t, w_1 is the age at seroconversion, and w_2 is the indicator variable, whose value is 1 if the time of seroconversion is after 1995, otherwise 0. The parameters $\gamma_l, \gamma_{1l}, \gamma_{2l}$, which specify the association between the marker and the survival process, allow to model the effect of the biomarker, adjusted for other covariates, on the competing events. We compare the parameters' estimates obtained for the longitudinal model in table 4.6.

The two models estimate the individual CD4 cell count pattern similarly, while they differ in estimating the population mean, as shown in figure 4.8. In figure 4.9, we represent the mean CD4 cell count pattern for each group.

Ignoring informative drop-outs for longitudinal process may lead to overoptimistic statements on marker trend, when patients in poorer health are more likely to leave the study, or may lead to underoptimistic statements, when patients in better health are more likely to leave the study. In accordance to this last statement, CD4 cell

Table 4.6. Coefficients estimates (CI(95%)) by piecewise linear mixed effects model

	Two-Stages		Joint	
α_1	22.280	(21.360, 23.200)	22.390	(21.750, 23.070)
α_2	0.756	(-0.340, 1.820)	0.542	(-0.334, 1.408)
α_3	2.502	(1.490, 3.532)	2.403	(1.564, 3.098)
α_4	-0.419	(-0.510, -0.327)	-0.449	(-0.482, -0.408)
α_5	0.077	(-0.032, 0.191)	0.107	(0.034, 0.168)
α_6	0.212	(0.101, 0.326)	0.253	(0.195, 0.305)
α_7	1.842	(1.491, 2.191)	1.847	(1.783, 1.939)
α_8	0.215	(-0.186, 0.623)	0.289	(0.113, 0.466)
α_9	-0.171	(-0.567, 0.229)	-0.247	(-0.396, -0.124)
α_{10}	-1.162	(-1.540, -0.787)	-1.144	(-1.229, -1.048)
α_{11}	-0.349	(-0.770, 0.078)	-0.434	(-0.589, -0.221)
α_{12}	-0.171	(-0.606, 0.251)	-0.081	(-0.188, 0.057)

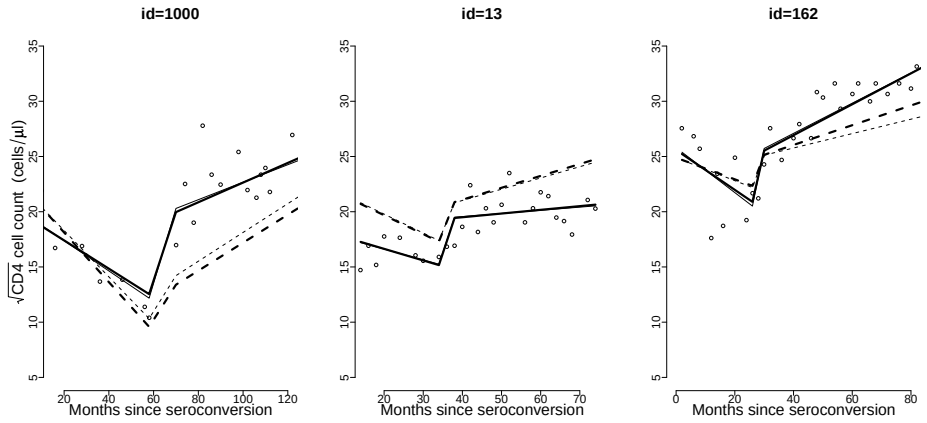

Fig. 4.8. By fitting the two-stages model (thin lines) and the joint model (thick lines), predicted individual (solid line) and mean (dash line) CD4 cell count pattern over time since seroconversion, for three individuals selected randomly, whose CD4 cell count is lower than 200 cell/μl, is included in [200, 350) cell/μl, and is higher than 350 cell/μl at HAART initiation, respectively.

count is lightly overestimated for individuals whose CD4 cell count is lower than 200 cell/μl at HAART initiation, while it is underestimated for those whose CD4 cell count is higher than 350 cell/μl. Similarly to model 4.1, the covariance matrix of random effects is given

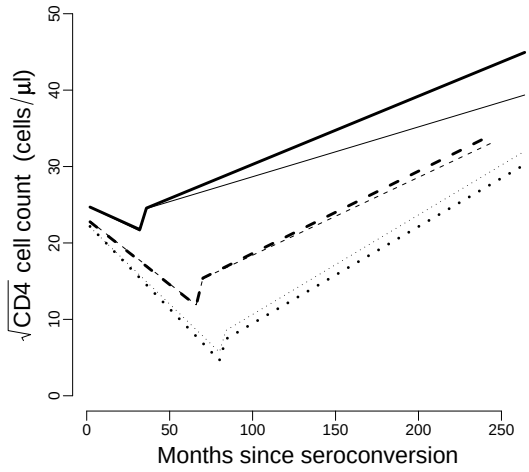

Fig. 4.9. By fitting the two-stages model (thin lines) and the joint model (thick lines), predicted mean CD4 cell count pattern over time since seroconversion, for three individuals selected randomly, whose CD4 cell count is lower than 200 cell/μl (solid line), is included in [200, 350) cell/μl (broken line), and is higher than 350 cell/μl (dotted line) at HAART initiation, respectively.

$$\begin{pmatrix} 22.880 & -0.407 & -1.835 & 2.259 \\ -0.407 & 0.219 & -0.119 & 0.020 \\ -1.835 & -0.119 & 2.173 & -2.005 \\ 2.259 & 0.020 & -2.005 & 2.312 \end{pmatrix}$$

In figure 4.10, observed CD4 cell count versus predicted CD4 cell count, and the residuals are represented.

In order to fit the survival model, we take normal priors for the parameters γ_l, γ_{1l}, γ_{2l}, ϕ_{1l}, and ϕ_{2l}, $l = 1, \ldots, 3$, and gamma priors for piecewise constant baseline hazards, $\lambda_{0l}(t) = \lambda_{jl}$, $t_{j-1} \leq t < t_j$, defined over some partitioning of the time scale into intervals not necessarily related to the times of covariate measurement. Our choice is based on the deciles of the observed time to each competing events. Yet the estimates do not change, varying the intervals of time. The coefficients' estimates are reported in table 4.7.

The coefficients' estimates, obtained by fitting two-stages and joint models, are very similar. The age at seroconversion does not seem to have a significant effect on the competing events. On the contrary, the individuals, whose date of seroconversion

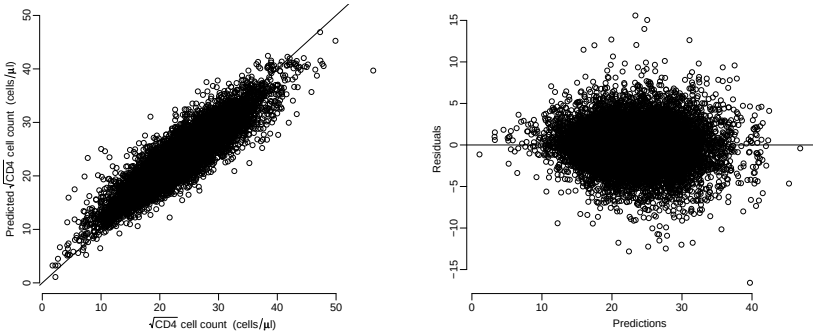

Fig. 4.10. First plot: observed CD4 cell count versus predicted CD4 cell count by joint modelling of CD4 cell count and survival data. Second plot: predicted CD4 cell count versus residuals.

Table 4.7. Coefficients' estimates (CI(95%)) by semiparametric proportional hazards model

	Two-stages		**Joint**	
ϕ_{11}	0.003	(-0.045, 0.043)	0.003	(-0.046, 0.004)
ϕ_{21}	1.293	(0.302, 2.445)	1.309	(0.306, 2.371)
γ_1	-0.117	(-0.245, 0.010)	-0.127	(-0.259, 0.004)
γ_{11}	-0.017	(-0.072, 0.049)	-0.015	(-0.077, 0.052)
γ_{21}	-0.005	(-0.068, 0.064)	-0.002	(-0.067, 0.071)
ϕ_{12}	-0.002	(-0.016, 0.011)	-0.002	(-0.017, 0.011)
ϕ_{22}	1.524	(1.179, 1.905)	1.513	(1.169, 1.873)
γ_2	0.093	(0.059, 0.128)	0.103	(0.061, 0.139)
γ_{12}	0.009	(-0.014, 0.033)	0.008	(-0.015, 0.034)
γ_{22}	0.021	(-0.0002, 0.044)	0.019	(-0.003, 0.046)
ϕ_{13}	0.007	(-0.005, 0.020)	0.007	(-0.005, 0.019)
ϕ_{23}	1.106	(0.809, 1.389)	1.106	(0.822, 1.390)
γ_3	0.059	(0.025, 0.094)	0.067	(0.035, 0.099)
γ_{13}	-0.007	(-0.024, 0.012)	-0.008	(-0.025, 0.010)
γ_{23}	-0.0005	(-0.019, 0.018)	-0.003	(-0.019, 0.015)

falls after 1995, have an higher probability of getting AIDS, of interrupting and changing therapy, than those whose date of seroconversion is before 1995. This result could be due to choice to performe the analysis over time since seroconversion. It might be that the individuals, who have seroconverted before 1995 and who have

started HAART as first therapy after 1995, are in better health than those who have seroconverted after 1995 and who have started the therapy immediately. The higher the CD4 cell count is, the lower the risk to fail from AIDS and the higher the probability to change the therapy and mainly to interrupt the therapy is. The failure hazards are not significantly different for the groups defined by a different CD4 cell count at HAART initiation.
Furthermore, we have fitted two further models, respectively given by

$$\lambda_l(t) = \lambda_{0l}(t) exp\{(\gamma_l + \gamma_{1l} I_1 + \gamma_{2l} I_2) u^{'}(t)\} \tag{4.4}$$

$$\lambda_l(t) = \lambda_{0l}(t) exp\{(\gamma_l + \gamma_{1l} I_1 + \gamma_{2l} I_2) u^{'}(t) + \phi_{1l} w_1\} \tag{4.5}$$

We have compared the models 4.2, 4.3, 4.4, by using the Deviance Information Criterion (DIC)(Spiegelhalter *et al.*, 2002). Thinking of β and y as the entire set of model parameters and data, $DIC = E[D(\beta|y)] + \{E[D(\beta|y)] - D(E[\beta|y])\} = \bar{D}(\beta) + p_D$, where $\bar{D}(\beta) = E_{\beta|y}[-2logf(y|\beta)] + 2logh(y)$, and p_D is the effective number of parameters. $f(y|\beta)$ is the likelihood function and $h(y)$ is some standardizing function of data alone. The fit of a model is summarized in the first term by the posterior expectation of the deviance function, $E[D(\beta|y)]$, while the complexity of the model si captured in the second term by the effective number of parameters p_D. We report the DIC of three fitted models in table 4.8. DIC1 is the component of DIC for the

Table 4.8. Deviance Information Criterion

Model	DIC1	DIC2	DIC3	DIC4	Dbar	p_D	DIC
cd4	449.16	3640.96	4346.94	74502.90	80202.70	2737.32	82940.00
cd4+age	450.42	3642.15	4341.34	74481.10	80168.00	2747.02	82915.00
cd4+age+calendar	445.74	3556.59	4281.54	74487.60	80021.80	2749.72	82771.50

longitudinal submodel, DIC2, DIC3, DIC4 the components for the survival submodel respectively for the risk 1,2, and 3. The DIC is the sum of three components. Based on DIC, the best model for survival process is the model 4.2.
We represent the hazard functions for three "average" subject in figure 4.11. By "average" subject, it is meant an individual, whose CD4 cell count pattern represents the population mean pattern over time since seroconversion and whose time of HAART initiation is the mean time of HAART initiation of the subjects under study and failed from the same cause. The three subjects have a different CD4 cell count at HAART initiation, < 200, [200, 350), and ≥ 350, while equal age at seroconversion and time of seroconversion. The variables effect on hazard function

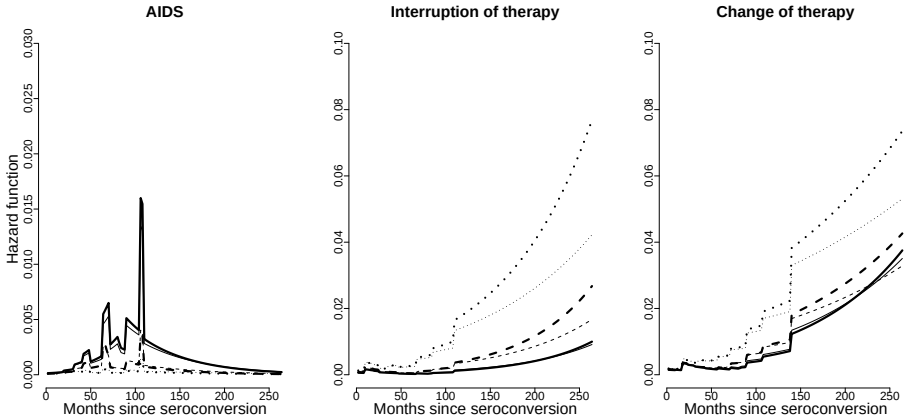

Fig. 4.11. Estimated hazard functions of three "mean" subjects, whose CD4 cell count is lower than 200 cell/μl (solid line), is included between 200 and 350 cell/μl (broken line), and is higher than 350 cell/μl (dotted line) at HAART initiation. The functions represent the istantaneous risk to fail from AIDS, to interrupt the therapy, and to change the therapy at time t, respectively.

could be different from that on cumulative incidence function, quantity of interest in a competing risks setting. Although we are not able to estimate it directly by fitting the Cox model, we can calculate the probability to fail from a competing risk, as first event, and represent it for a "mean" subject, as shown in figure 4.12. The estimated cumulative incidence curves by two-stages and joint models are very similar, due to similarity of hazard functions. The individuals whose CD4 cell count is lower than 200 cell/μl have an higher probability to fail from AIDS and a lower probability to interrupt the therapy than the individuals whose CD4 cell count at HAART initiation is included between 200 and 350 cell/μl and those whose CD4 cell count is higher than 350 cell/μl. The individuals whose CD4 cell count is included between 200 and 350 cell/μl at HAART initiation have an higher probability to fail from AIDS and a lower probability to interrupt the therapy than those whose CD4 cell count is higher than 350 cell/μl. The probability to change therapy does not seem to differ between groups. Hence the CD4 cell count is an important predictor for the time to competing events. By adjusting the analysis for time since seroconversion, CD4 cell count pattern and age at seroconversion, we have evaluated the risk to fail from competing risks for three groups defined according to CD4 cell count at HAART initiation, and it seems better to start the therapy when the biomarker value is higher than 350 cell/μl.

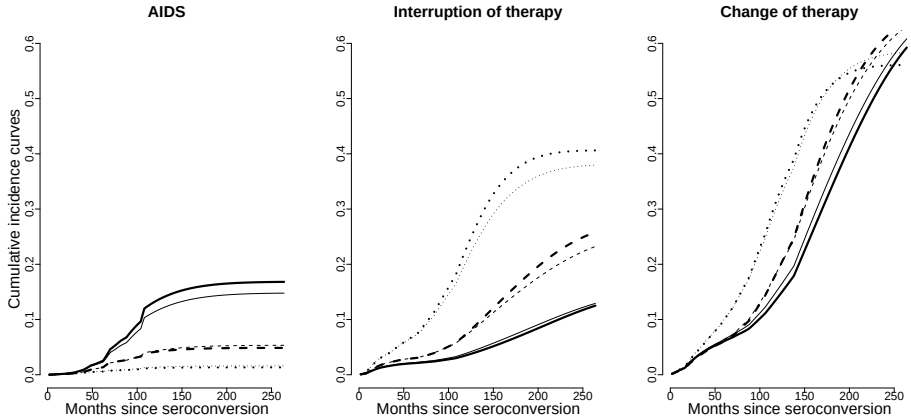

Fig. 4.12. Estimated cumulative incidence functions of three "mean" subjects, whose CD4 cell count is lower than 200 cell/μl (solid line), is included between 200 and 350 cell/μl (broken line), and is higher than 350 cell/μl (dotted line) at HAART initiation. The functions represent the risk to fail from AIDS, to interrupt the therapy, and to change the therapy, as first event, within time t, respectively.

4.2.2 Bivariate longitudinal model and competing risk

Likewise to CD4 cell count analysis, by using the deciles of distribution of variable "elapsed time between seroconversion and HAART" as stratification criteria, we represent the box-plots of viral load for each interval of time equal to two months and a smoothing function, which uses locally-weighted polynomial regression, remarking the interval of time the individuals start the treatment. We show the viral load pattern of individuals who start treatment between 94 and 122 months after seroconversion in figure 4.13. Analogue plots are obtained for each strata. Approximately the biomarker increases after seroconversion, reaching its maximum value before the initiation of HAART, it decreases very rapidly in the first months after submission to therapy, and then it tends to stabilize. After choosing an interval of time for start of treatment arbitrarily, we represent the plots of viral load for the individuals whose viral load is lower than 10000 copies/mL, and those whose viral load is higher than 10000 copies/mL at HAART initiation, as shown in figure 4.14. The higher the viral load is at seroconversion, the higher it is at HAART initiation, and the more effective the therapy seems to be. Generally the two groups show a similar viral load pattern. We indicate by $\{y_{ij}^{''} : j = 1, \ldots, n_i^{''}\}$ the viral load measurements at times $\{t_{ij} : j = 1, \ldots, n_i^{''}\}$. Since our aim is to model the viral load and CD4 cell count and their relation by evaluating the change points, we fit the

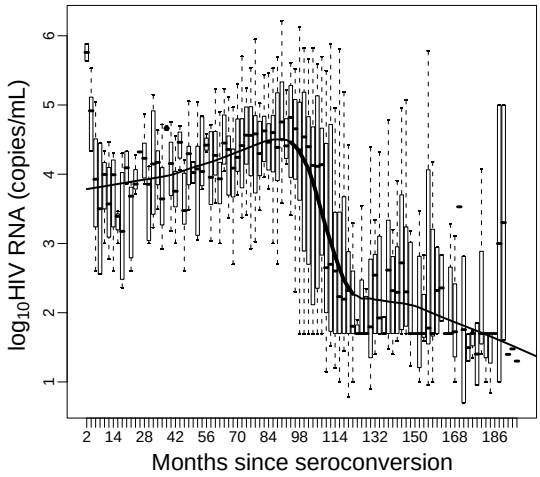

Fig. 4.13. Viral load pattern of individuals, who start therapy between 94 and 122 months after seroconversion

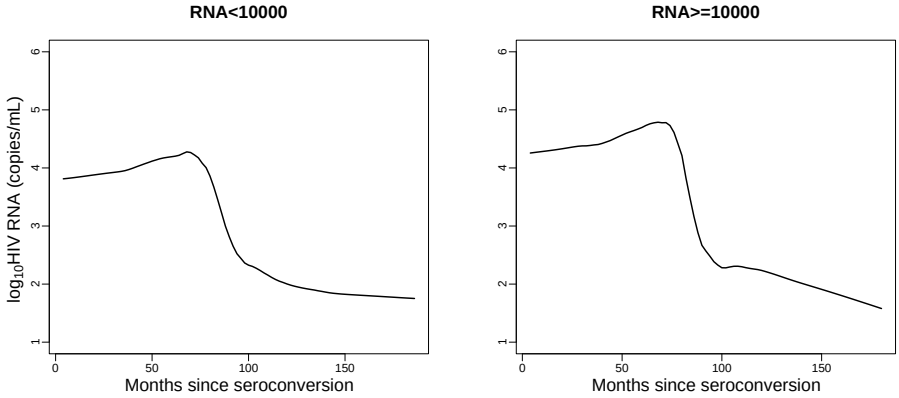

Fig. 4.14. Viral load pattern of each group, who started therapy between 76 and 94 months after seroconversion and characterized by different viral load at HAART initiation, <10000, ≥ 10000 respectively.

two biomarkers' pattern by a bivariate piecewise linear mixed effects model, given by

$$\begin{cases} y_{ij}^{'} = u_{ij}^{'} + \epsilon_{ij}^{'} \\ y_{ij}^{''} = u_{ij}^{''} + \epsilon_{ij} \end{cases} \tag{4.6}$$

where $u_{ij}^{'}$ is defined in model 4.1, and $u_{ij}^{''}$ is defined as

$$\begin{aligned} u_{ij}^{''} = {} & \alpha_1^{'} + \alpha_2^{'} I_3 + (\alpha_3^{'} + \alpha_4^{'} I_3) t_{ij} + (\alpha_5^{'} + \alpha_6^{'} I_3)(t_{ij} - t_1) I(t_{ij} - t_1 > 0) \\ & + (\alpha_7^{'} + \alpha_8^{'} I_3)(t_{ij} - t_2) I(t_{ij} - t_2 > 0) + \theta_{5i} + \theta_{6i} t_{ij} + \theta_{7i}(t_{ij} - t_1) \\ & I(t_{ij} - t_1 > 0) + \theta_{8i}(t_{ij} - t_2) I(t_{ij} - t_2 > 0). \end{aligned}$$

I_3 is an indicator variable, having value 1 if viral load is higher than 10000 copies/mL at the therapy initiation, 0 otherwise, t_{i1} is the time of HAART initiation dependent on ith subject, and t_{i2} is the time when the slope changes because of therapeutic effect's decrease, dependent on ith subject. Equally to model 4.1, we fix t_{i2} as the time after three months since HAART initiation, to evaluate the changes in viral load pattern parallely to those of CD4 cell count. $\boldsymbol{\alpha}^{'}$ is a 8×1 vector of unknown fixed effects, $\boldsymbol{\theta}_i$ is a 8×1 vector of unobservable random effects, and $\boldsymbol{\epsilon}_i^{''}$ is a within-individual residuals vector. The random effects $\boldsymbol{\theta}_i$ are assumed to be normally distributed with mean $\mathbf{0}$ and 8×8 variance-covariance matrix $\boldsymbol{\Sigma}$. The $\boldsymbol{\theta}_i$ are distributed indipendently of each other and of the within-subjects residuals $\epsilon_{ij}^{'}$ and $\epsilon_{ij}^{''}$. The $\boldsymbol{\epsilon_i^{''}}$ are assumed to be indipendent and normally distributed with mean $\mathbf{0}$ and $n_i^{''} \times n_i^{''}$ variance-covariance matrix $\sigma_{\epsilon^{''}}^2 I_{n_i^{''}}$, where $I_{n^{''}}$ denotes the $n^{''} \times n^{''}$ identity matrix. It is known that the viral load and CD4 cell count are negatively correlated, yet their relationship may not be costant in the time and may depend on the subject. We model the dependence between the viral load and CD4 cell count by the covariance matrix the individual effects $\boldsymbol{\theta_i}$, Σ. We take multivariate normal priors for the main effects vector $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{'}$, an inverse gamma priors for the error variance $\sigma_{\epsilon^{'}}^2$ and $\sigma_{\epsilon^{''}}^2$, and an inverse Wishart for the common parameter Σ, all having very low precision. We monitor the MCMC convergence by three parallel MCMC sampling chains of 80000 iterations each, following a 30000-iteration "burn-in" period. By selecting two subjects randomly, we show the individual deviation from overall effect, due to the inclusion of random effects, in figure 4.15.

The viral load pattern of two subjects differs mainly in the period before HAART initiation, when the viral load of the first individual is almost constant from sero-conversion to initiation of the therapy, while is increasing for the second individual. The coefficients' estimates obtained by fitting univariate and bivariate models for CD4 cell count and viral load is reported in table 4.9. By univariate models we mean the models 4.1 and 4.4 for CD4 cell count and viral load respectively, omitting the dependence between the two biomarkers expressed by Σ.

In figure 4.16 and 4.17 the CD4 cell count and viral load patterns are represented,

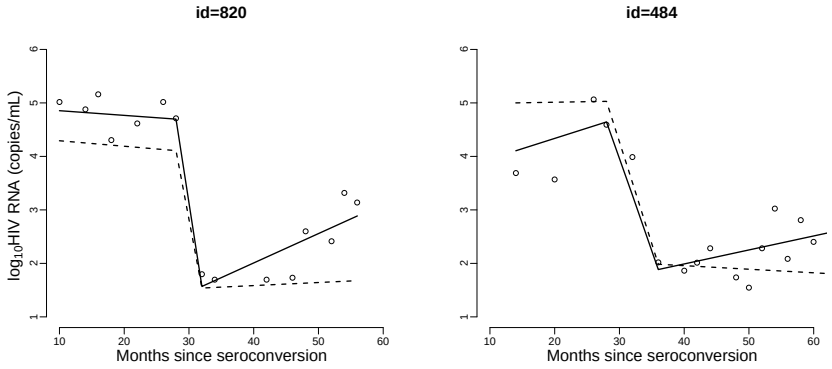

Fig. 4.15. Predicted individual (solid line) and mean (dash line) viral load pattern over the time since seroconversion for two individuals selected randomly, whose viral load is lower than 10000 copies/mL, and is higher than 10000 copies/mL respectively.

for both models.

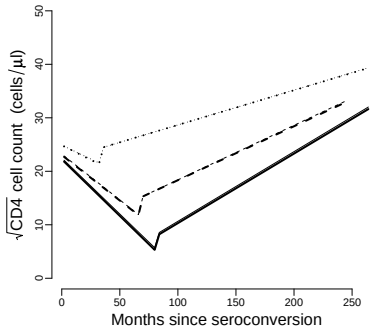

Fig. 4.16. Comparison between the mean CD4 cell count pattern over the time since seroconversion estimated by longitudinal univariate (thin line) and bivariate (thick line) models respectively, when CD4 cell count is lower than 200 cell/μl (solid line), is included between 200 and 350 cell/μl (broken line), and is higher than 350 cell/μl (dotted line) respectively.

The mean CD4 cell count pattern does not differ between the two models, while the estimated mean viral load by bivariate model is lower than that estimated by

Table 4.9. Coefficients' estimates (CI(95%)) of longitudinal univariate and bivariate models

	Bayesian univariate		**Bayesian bivariate**	
α_1	22.280	(21.360, 23.200)	22.020	(21.110, 22.950)
α_2	0.756	(-0.340, 1.820)	0.844	(-0.214, 1.872)
α_3	2.502	(1.490, 3.532)	2.745	(1.763, 3.716)
α_4	-0.419	(-0.510, -0.327)	-0.423	(-0.505, -0.334)
α_5	0.077	(-0.032, 0.191)	0.083	(-0.021, 0.183)
α_6	0.212	(0.101, 0.326)	0.203	(0.101, 0.298)
α_7	1.842	(1.491, 2.191)	1.881	(1.523, 2.230)
α_8	0.215	(-0.186, 0.623)	0.211	(-0.193, 0.624)
α_9	-0.171	(-0.567, 0.229)	-0.090	(-0.495, 0.320)
α_{10}	-1.162	(-1.540, -0.787)	-1.198	(-1.569, -0.829)
α_{11}	-0.349	(-0.770, 0.078)	-0.352	(-0.783, 0.078)
α_{12}	-0.171	(-0.606, 0.251)	-0.243	(-0.672, 0.182)
$\alpha_1^{'}$	4.385	(4.268, 4.502)	4.416	(4.297, 4.540)
$\alpha_2^{'}$	0.589	(0.414, 0.760)	0.536	(0.372, 0.698)
$\alpha_3^{'}$	-0.021	(-0.044, -0.001)	-0.032	(-0.071, 0.002)
$\alpha_4^{'}$	0.025	(-0.021, 0.068)	-0.015	(-0.027, 0.053)
$\alpha_5^{'}$	-1.264	(-1.322, -1.204)	-1.275	(-1.338, -1.212)
$\alpha_6^{'}$	-0.258	(-0.350, -0.171)	-0.242	(-0.332, -0.153)
$\alpha_7^{'}$	1.296	(1.232, 1.360)	1.315	(1.254, 1.377)
$\alpha_8^{'}$	0.208	(0.110, 0.2089)	0.193	(0.102, 0.286)

univariate model. The figure 4.17 suggests that the therapy has more effect on the patients, whose viral load is higher than 10000 copies/mL than those, whose viral load is lower than 10000 copies/mL at HAART initiation. This result may be due to heterogeneity inside the group of individuals, whose viral load is higher than 10000 copies/mL. It would be correct to create more groups, defined according to viral load at HAART initiation, i.e. <10000, [10000,100000), and $\geq$ 10000. Furthermore the bivariate model allows to estimate the correlation matrix between the individual slopes of each marker, given by

$$\begin{pmatrix} -0.952 & 0.814 & -0.059 & -0.350 \\ 0.693 & -0.610 & 0.219 & 0.098 \\ 0.757 & -0.651 & -0.307 & 0.581 \\ -0.794 & 0.679 & 0.273 & -0.576 \end{pmatrix}$$

As expected, CD4 cell count and viral load are negatively correlated: $\rho = -0.952$, std=0.26 for the intercept, $\rho = -0.610$, std=0.01 for the first slope, $\rho = -0.307$, std=0.04 for the third slope, and $\rho = -0.576$, std=0.05 for the last slope.

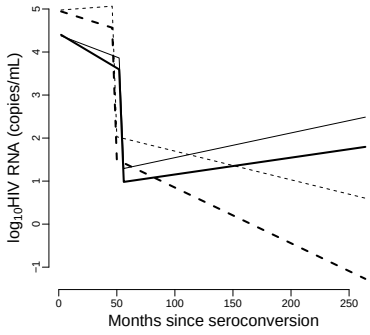

Fig. 4.17. Comparison between the mean (broken line) viral load pattern over the time since seroconversion estimated by longitudinal univariate (thin line) and bivariate (thick line) models respectively, when viral load is lower than 10000 copies/mL (solid line), and is higher than 10000 copies/mL (dotted line) respectively.

In order to evaluate the joint variation over time of the two biomarkers and of competing events we model jointly the longitudinal and survival process. Therefore, we modify the model 4.2, by including both biomarkers, CD4 cell count and viral load as follows

$$\lambda_l(t) = \lambda_{0l}(t) exp\{(\gamma_l + \gamma_{1l} I_1 + \gamma_{2l} I_2) u^{'}(t) + (\beta_l + \beta_{1l} I_3) u^{''}(t) + \phi_{1l} w_1 + \phi_{2l} w_2\}$$

where $\lambda_{0l}(t)$ is the baseline hazard function at time t for failure l and $I_3 = 1$ if viral load is higher than 10000 copies/mL, otherwise 0. We take normal priors for the parameters γ_l, γ_{1l}, γ_{2l}, β_l, β_{1l}, ϕ_{1l}, ϕ_{2l}, and gamma priors for piecewise costant baseline hazards $\lambda_{0l}(t)$, all having very low precision. We report the results obtained by fitting the two biomarkers pattern by separate univariate models, two-stages model, and finally joint model, in table 4.10, and we compare the population mean CD4 cell count and viral load pattern in figure 4.18 and 4.19 respectively. The CD4 cell count pattern is similar to that obtained by fitting models 4.1 and 4.2 jointly, consistently to similarity of CD4 cell pattern by fitting univariate (4.1) and bivariate models (4.5). Without considering informative dropouts, viral load pattern is overestimated or underestimated, depending on the health status of the subject, who leaves the study. In figure 4.20, observed versus predicted CD4 cell count, and residuals are represented. Analogue plots are shown in figure 4.21 for the viral load. Now we focus on survival analysis, comparing the coefficients' estimates obtained by two-stages and joint model in table 4.11. For clarity, we represent only the cumulative incidence curves for the three competing events, obtained by

Table 4.10. Coefficients' estimates (CI(95%)) of separate, two-stages, and joint models for longitudinal data

	Separate		**Two-stages**		**Joint**	
α_1	22.280	(21.360, 23.200)	22.020	(21.110, 22.950)	22.010	(21.040, 23.050)
α_2	0.756	(-0.340, 1.820)	0.844	(-0.214, 1.872)	0.820	(-0.346, 1.857)
α_3	2.502	(1.490, 3.532)	2.745	(1.763, 3.716)	2.788	(1.654, 3.938)
α_4	-0.419	(-0.510, -0.327)	-0.423	(-0.505, -0.334)	-0.420	(-0.502,-0.338)
α_5	0.077	(-0.032, 0.191)	0.083	(-0.021, 0.183)	0.080	(-0.011,0.168)
α_6	0.212	(0.101, 0.326)	0.203	(0.101, 0.298)	0.205	(0.107, 0.303)
α_7	1.842	(1.491, 2.191)	1.881	(1.523, 2.230)	1.869	(1.729, 2.007)
α_8	0.215	(-0.186, 0.623)	0.211	(-0.193, 0.624)	0.246	(-0.014, 0.473)
α_9	-0.171	(-0.567, 0.229)	-0.090	(-0.495, 0.320)	-0.080	(-0.348, 0.270)
α_{10}	-1.162	(-1.540, -0.787)	-1.198	(-1.569, -0.829)	-1.215	(-1.337, -0.998)
α_{11}	-0.349	(-0.770, 0.078)	-0.352	(-0.783, 0.078)	-0.355	(-0.586, -0.073)
α_{12}	-0.171	(-0.606, 0.251)	-0.243	(-0.672, 0.182)	-0.180	(-0.592, 0.231)
$\alpha_1^{'}$	4.385	(4.268, 4.502)	4.416	(4.297, 4.540)	4.359	(4.227, 4.475)
$\alpha_2^{'}$	0.589	(0.414, 0.760)	0.536	(0.372, 0.698)	0.521	(0.340, 0.702)
$\alpha_3^{'}$	-0.021	(-0.044, -0.001)	-0.032	(-0.071, 0.002)	-0.028	(-0.039, -0.007)
$\alpha_4^{'}$	0.025	(-0.021, 0.068)	-0.015	(-0.027, 0.053)	0.012	(-0.032, 0.050)
$\alpha_5^{'}$	-1.264	(-1.322, -1.204)	-1.275	(-1.338, -1.212)	-1.212	(-1.254, -1.177)
$\alpha_6^{'}$	-0.258	(-0.350, -0.171)	-0.242	(-0.332, -0.153)	-0.373	(-0.439, -0.313)
$\alpha_7^{'}$	1.296	(1.232, 1.360)	1.315	(1.254, 1.377)	1.260	(1.216, 1.303)
$\alpha_8^{'}$	0.208	(0.110, 0.289)	0.193	(0.102, 0.286)	0.322	(0.262, 0.384)

fitting the joint model, and compare the individuals with different CD4 cell count and viral load at HAART initiation, in figure 4.22. The viral load is an important predictor of the probability of getting AIDS, while it is less important in determining the interruption and the modification of the therapy. Adjusting the analysis for viral load does not lead to coefficients' estimates for CD4 cell count very different from those obtained by modelling jointly the models 4.1 and 4.2, yet it allows for differentiating the cumulative incidence curves according to viral load at HAART initiation. By comparing the figures 4.12 and 4.22, one can note that the probability of failing from AIDS is a "mean" of the same probabilities of patients whose viral load is lower and higher than 10000 copies/mL at initiation of therapy, respectively. For the six groups, defined according CD4 cell count and viral load at HAART initiation, it is visible the unequal elapsed time between seroconversion and start of treatment. If we had considered the time since HAART initiation as time-scale, we would have observed all cumulative incidence curves relocated at time of HAART initiation of the patients whose CD4 cell count is lower than 200 cell/μl (between 50

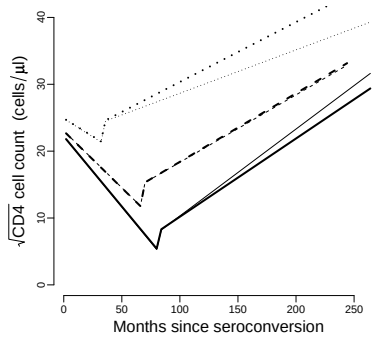

Fig. 4.18. Mean CD4 cell count estimated by a bivariate linear mixed effects model (thin line) and by a joint model to viral load and competing events (thick line), for subjects whose CD4 cell count is lower than 200 cell/μl (solid lines), is included between 200 and 350 cell/μl (broken lines), and is higher than 350 cell/μl (dotted line).

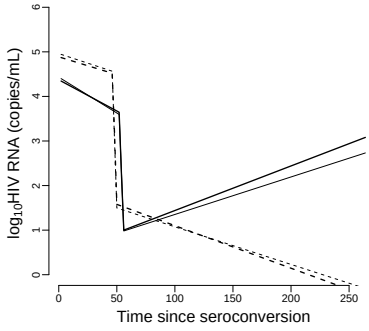

Fig. 4.19. Mean viral load estimated by a bivariate linear mixed effects model (thin line) and by a joint model to CD4 cell count and competing events (thick line), for subjects whose viral load is lower than 10000 copies/mL (solid lines), and is higher than 10000 copies/mL (broken line).

and 100 months after seroconversion), the distance between the curves would have been bigger, and the estimate would have been biased. Furthermore it is evident how the cumulative incidence curves compensate each other, i.e. individuals, whose CD4 cell count and viral load are lower than 200 cell/μl and higher than 10000 copies/mL (in bad health) at HAART initiation, have an higher probability of

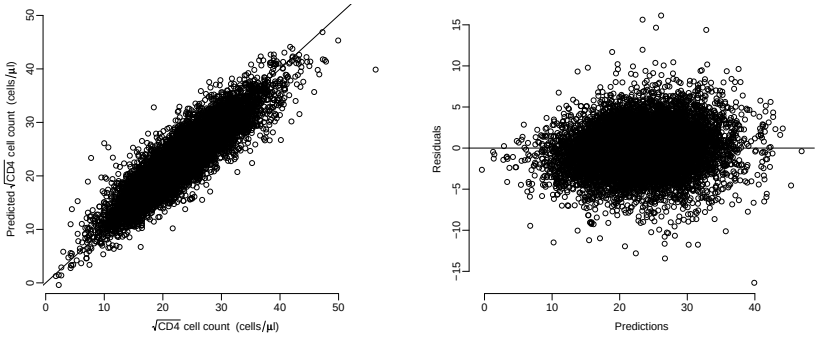

Fig. 4.20. First plot: observed CD4 cell count versus predicted CD4 cell count by joint modelling of CD4 cell count, viral load and survival data. Second plot: predicted CD4 cell count versus residuals.

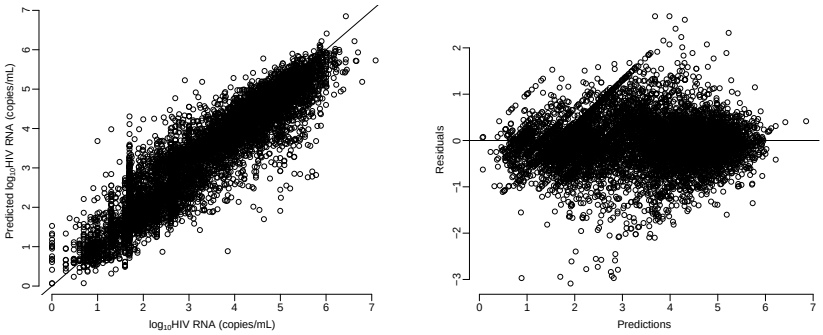

Fig. 4.21. First plot: observed viral load versus predicted viral load by joint modelling of CD4 cell count, viral load and survival data. Second plot: predicted viral load versus residuals.

getting AIDS and at the same time a lower probability of changing therapy than individuals, whose CD4 cell count and viral load are both lower than 200 cell/μl and 10000 copies/mL, respectively. Briefly it summarizes the main characteristic of cumulative incidence function in a competing risks setting.

For simplicity, some important aspects have been omitted in modelling the viral load pattern, i.e. the left censoring because of quantification limit of assay used to

Table 4.11. Coefficients' estimates (CI(95%)) by semiparametric proportional hazards model

	Two-stages		Joint	
ϕ_1	-0.004	(-0.055, 0.042)	0.000	(-0.049, 0.044)
ϕ_2	-0.008	(-0.023, 0.006)	-0.005	(-0.019, 0.009)
ϕ_3	0.000	(-0.014, 0.012)	0.006	(-0.007, 0.019)
ϕ_1	1.045	(0.000, 2.181)	1.225	(0.168, 2.338)
ϕ_2	1.019	(0.653, 1.392)	1.135	(0.739, 1.530)
ϕ_3	0.511	(0.205, 0.845)	0.666	(0.328, 0.985)
γ_1	-0.130	(-0.269, -0.004)	-0.146	(-0.283, -0.016)
γ_{11}	-0.015	(-0.074, 0.055)	-0.004	(-0.067, 0.073)
γ_{21}	0.000	(-0.065, 0.076)	0.019	(-0.051, 0.096)
γ_2	0.052	(0.011, 0.091)	0.082	(0.041, 0.123)
γ_{12}	0.008	(-0.017, 0.036)	0.002	(-0.026, 0.028)
γ_{22}	0.021	(-0.003, 0.051)	0.011	(-0.017, 0.037)
γ_3	0.013	(-0.021, 0.046)	0.039	(0.003, 0.070)
γ_{13}	-0.005	(-0.022, 0.012)	-0.010	(-0.028, 0.010)
γ_{23}	0.002	(-0.016, 0.020)	-0.004	(-0.022, 0.017)
β_1	-0.510	(-1.035, -0.092)	-0.425	(-0.938, 0.050)
β_{11}	0.488	(0.174, 0.880)	0.461	(0.112, 0.852)
β_2	-0.581	(-0.703, -0.452)	-0.800	(-0.938, -0.660)
β_{21}	0.049	(-0.050, 0.154)	0.219	(0.110, 0.332)
β_3	-0.618	(-0.740, -0.502)	-0.792	(-0.926, -0.669)
β_{31}	-0.008	(-0.102, 0.081)	0.218	(0.13, 0.324)

measure it, how it appears in figure 4.21. It would be appropriate to repeat the previous analysis extended to viral load, also considering those features.

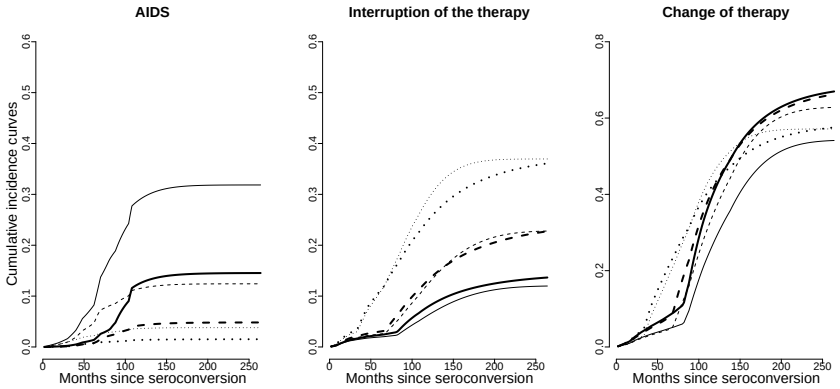

Fig. 4.22. Comparison of cumulative incidence curves of the three competing events, for the groups defined according their CD4 cell count and viral at HAART initiation: CD4$<$ 200 (solid line), CD4 in $[200, 350)$ (broken line), CD4 $\geq$ 350 (dotted line); RNA$<$ 10000 (thick line), RNA$\geq$ 10000 (thin line).

5

Conclusions

5.1 Discussion

Motivated by an observational study on HIV data, provided by CASCADE, we have presented an extension of joint modelling of longitudinal and survival data to competing risks framework. The most medical-epidemiological studies are characterized by both repeated measurements over time and multiple events, that cause the patient's exit from the study, i.e. death and recurrence of a disease. The response outcomes of longitudinal process may be important predictors to survival process, so as the survival process may be informative for the longitudinal process. Very often the measurements are taken on at irregular times, in unequal number for each patient, and may be prone to measurement error. Hence we need to model the longitudinal process in order to evaluate its tie with the survival process. At the same time, the survival process may auto-select a part of the starting sample over time, by the occurrence of certain events, generating informative dropouts for the longitudinal process. For istance the patients, who are in poorer health, will have an higher probability of leaving the study than those in good health, and the estimate of longitudinal process will be based only on a share of sample, resulting in such a way biased. The joint modelling of longitudinal and survival data allows to adjust the longitudinal analysis for informative dropouts and to utilize the "true" response outcomes to predict the time to occurrence of event of interest. Furthermore, it makes a more efficient use of the data, because both longitudinal and survival informations are used at the same time to obtain the parameters' estimates of the underlying model. When the aim is to study the time elapsed from some particular starting point to the occurrence of an event in presence of several possible events, the extension to competing risks framework is required. A competing event is an event whose occurrence either precludes the occurrence of another events under examination or alters the probability of occurrence of these other

events. Performing the survival analysis by classical method, i.e. by Kaplan-Meier curves, would produce biased results in presence of competing events.
Several approaches have been proposed to model jointly repeated measurements and the time to an event, by both classical and Bayesian methods (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997; Henderson *et al.*, 2000; Wang and Taylor, 2001; Guo and Carlin, 2004; Berzuini and Allemani, 2004). We have moved from these methods to model jointly longitudinal data and time to competing events. Specifically, we have modelled the longitudinal process by a linear mixed effects model and the time to competing risks by a proportional hazards model. We have explained how to develop the EM algorithm, by exploiting a property of partial likelihood of Cox model, and how to deduce the full conditional distributions, by a Bayesian approach.
We have applied our methodology to an observational study regarding progression of HIV to AIDS. Given increasing concern regarding the challenges of maintaining HAART regimens, determining the optimal time when to initiate the therapy is of clinical interest. By using data collected by CASCADE, we have evaluated the probability of getting AIDS, of interrupting and of changing therapy for the men, who has got HIV by homosexual relations and whose CD4 cell count is lower than 200 cell/μl, is included between 200 and 350 cell/μl, and finally is higher than 350 cell/μl at HAART initiation, respectively. Considering the interruption and the change of therapy as competing events allows to solve out the problem of dependent censoring for the occurrence of AIDS, and then to estimate its probability correctly. Since the time of seroconversion is reliably estimated and data on CD4 cell count and viral load are collected, we model the two biomarkers over time since seroconversion, and we control the survival analysis for bias due to markers pattern before HAART initiation. As shown by previous studies, CD4 cell count and viral laod are negatively correlated, and the individuals, whose CD4 cell count is lower than 200 cell/μl at HAART initiation have an higher probability of getting AIDS, while a lower probability of interrupting the therapy. By a comparison of the three groups of patients, it appears that those whose CD4 cell count is included between 200 and 350 cell/μl at HAART initiation has an higher probability of getting AIDS, and a lower probability of interrupting the therapy, than those whose CD4 cell count is higher than 350 cell/μl. As expected, the patients who start the therapy when the CD4 cell count is lower than 200 cell/μl, have an higher probability of getting AIDS. Therefore the risk of AIDS differs between those delaying treatment until their CD4 cell count is 200-349 cell/μl and those who initiate therapy with CD4 cell count $\geq$ 350 cell/μl.
Furthermore, in performing the analysis, we compare the results obtained by first

modelling the longitudinal data and then substituting the fitted values in survival model and by joint modelling. Although the cumulative incidence curves estimated by the two approaches are similar, some differences are visible in the longitudinal analysis. CD4 cell count and viral load are overestimated or underestimated by the first approach, since informative dropouts are not considered.
Actually we have used a particular specification of a joint model that is appropriate to the dataset that we have analyzed. Yet, the approach is generalizable to many different situations, characterized by both longitudinal and survival processes in the presence of competing events, such as cancer studies where the effect of an allocated therapy on the patient's antibody levels against tumor cells is modelled jointly with a competing events survival process.

5.2 Further research

Because of complexity of data, we have neglected some important aspects, that we will deal with in the near future.
We have not taken into account the quantification limit of the assays used to quantify viral load. After the HAART initiation a lot of patients experience a drastic fall of viral load below assays quantification limit. Approaches, like imputation of half the limit of the assay threshold may lead to biased estimation of model parameters and their standard errors (Jacqmin-Gadda *et al.*, 2000). When one or more markers present left-censored values, the likelihood needs to be modified. For ith subject, let $\boldsymbol{y}_i^0 = (y_{i1}^0, y_{i2}^0, \ldots, y_{in_{i0}}^0)$ the vector of observed response variable, $\boldsymbol{y}_i^c$ the vector of censored outcomes and $\boldsymbol{c}_i$ the n_i^c vector of measurement thresholds. Then the contribution of y_i to the conditional likelihood is given by

$$f(y_i)P(y_i^c < c_i) = \Big(\prod_{j=1}^{n_{i0}} f(y_{ij}^0)\Big)\Big(\prod_{j=1}^{n_{ic}} F(c_{ij})\Big)$$

where $f(y_{ic})$ and $F(y_{ic})$ are the probability density function and the cumulative distribution function of y_{ic}, respectively. Furthermore it would be appropriate to introduce an indicator variable of the assay type used to measure viral load in the longitudinal model for viral load.
We have considered two categories of patients, as regards to their viral load at HAART initiation, <10000 and ≥10000 copies/mL, that is the value corresponding the median of viral load at the initiation of therapy. Yet, in order to have more homogeneous groups, it would be better to split the champion in three groups, $<$ 10000, in (10000, 100000], and ≥100000 copies/mL. Since we are interested in expanding the method to competing risks framework, the problem is that we need

to observe the competing events in each category, and mainly for AIDS it is not always possible if the groups are too many.
Although our choice to model the longitudinal process by a piecewise linear mixed effects model is motivated by the search of simplicity in the interpretation of the parameters, it could fail fitting some shapes of biomarkers' pattern. Hence for example, the interest in modelling the longitudinal process by a model, that incorporates a mean structure dependent on covariates, a random intercept, a measurement error and an integrated Ornstein-Uhlenbeck stochastic process (Wang and Taylor, 2001). By including the IOU process, the longitudinal component of the model provides a more flexible and plausible structure for individual's marker pattern than do the standard random effects model. Indeed its parameters control the amount of smoothness of a person's path without imposing any particular shapes on the path, and it alloes for random effects and Brownian motion as special cases.
It would be also interesting to evaluate if modelling of autocorrelations through use of time series models for the errors, such as an autoregressive process of order one (AR(1)), in addition to the inclusion of random effects terms may be more appropriate and may lead to a more proper representation of correlation structure. Generalizations of the AR(1) model to other time series models, such as higher order AR model, may also be worth of consideration.
A further aspect to develop, is the modelling of the survival component by the Fine & Gray model (Fine and Gray, 1999). Indeed it would allow to estimate the effect of covariates of interest on the cumulative incidence functions directly, and not only on the hazards function of the competing events.
Finally, in order to make a comparison between the results, obtained by the frequentist and the Bayesian approaches, the development of the EM-based algorithm, described in chapter 3, is required.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, (6): 716–723.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Models based on Counting Processes*. Berlin, New York: Springer-Verlag.

Arriagada, R., Rutqvist, L., and Kramar, A. (1992). Competing risks determining event-free survival in early breast cancer. *British Journal of Cancer*, **66**, 951–957.

Bacchetti, P. and Moss, A. (1989). Incubation period of aids in san francisco. *Nature*, **338**, 251–253.

Bates, D. and Pinheiro, J. (1998). Computational methods for multilevel models. *Technical memorandum bl0112140-980226-01tm*, pages 1–29.

Berzuini, C. and Allemani, C. (2004). Effeciveness of potent antiretroviral therapy on progression of human immunodeficiency virus: Bayesian mdelling and model checking via counterfactual replicates. *Applied Statistics*, **53**, 633–650.

Berzuini, C. and Larizza, C. (1996). A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, **18**, No.2, 109–123.

Carlin, B. and Luis (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman Hall.

Carpenter, C., Cooper, D., Fischl, M., Gatell, J., Gazzard, B., Hammer, S., Hirsch, M., Jacobsen, D., Katzenstein, D., Montaner, J., Richman, D., Saag, M., Schechter, M., Schooley, R., Thompson, M., Vella, S., Yeni, P., and Volberding, P. (2000). Antiretroviral therapy in adults. updated recommendations of the international aids society-usa panel. *Journal of the American Medical Association*, **283**, 381–390.

Cole, S., Li, R., Anastos, K., Detels, R., Young, M., Chmiel, J., and Munoz, A. (2004). Accounting for leadtimein cohort studies: evaluating when to initiate hiv

therapies. *Statistics in medicine*, **23**, 3351–3363.

Coviello, V. and Boggess, M. (2004). Cumulative incidence estimation in the presence of competing risks. *The Stata Journal*, **4**, No. 2, 103–112.

Cox, D. (1972). Regression models and life tables. *Jornal of the Royal Statistical Society, Series B*, **34**, 187–220.

Cox, D. and Oakes, D. (1984). *Analysis of survival data*. Chapman and Hall: London.

Crowder, M. (2001). *Classical Competing Risks*. Chapman and Hall/CRC.

DeGruttola, V. and Tu, X. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

DeGruttola, V., Lange, N., and Dafni, U. (1991). Modelling the progression of hiv infection. *Journal of the American Statistical Association*, **86**, 569–577.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, (1): 1–38.

Dempster, A., Rubin, D., and Tsutakawa, R. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, **76**, 341–353.

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

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

Elashoff, R., Li, G., , and Li, N. (2007). An approach to joint analysis of ongitudinal measurements and competing risks failure time data. *Statistics in Medicine*, **26**, 2813–2835.

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

Fearn, T. (1975). A bayesian approach to growth curves. *Biometrika*, **62**, 89–100.

Fine, J. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics*, **2**, 1, 85–97.

Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**, 496–509.

Fleming, T. and Arrington, D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

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

Galai, N., Munoz, A., and Chen, K. (1993). Tracking of markers and onset of disease among hiv seroconverters. *Statistics in medicine*, **12**, 2133–2145.

Gallant, A. R. and Nichka, D. (1987). Seminonparamteric maximum likelihood estimation. *Econometrica*, **55**, 363–390.

Gaynor, J., Fener, F., Tan, C., Wu, D., Little, C., Straus, D., Clarkson, B., and Brennan, M. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from oncology data. *Journal of the Statistical Association*, **18**, 400–409.

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

Gooley, T., Leisenring, W., Crowley, J., and Storer, B. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in medicine*, **18**, 695–706.

Grant, L., Yamashita, T., Phair, J., Detels, R., Wolinsky, S., Margolick, J., Rinaldo, C., and Jacobson, L. (2003). When to initiate highly active antiretroviral therapy: a cohort approach. *American Journal of Epidemiology*, **157**, 738–746.

Gray, R. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*, **16**, 1141–1154.

Guo, X. and Carlin, B. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, No. 1, 1–9.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to relaed problems. *Journal of the American Statistical Association*, **72**, 320–340.

Hastings, W. (1970). Monte carlo sampling methods using markov chain and their application. *Biometrika*, **57**, 549–603.

Hayes, W. (1973). *Statistical for Social Sciences*. New York: Holt, Rinehart and Winston.

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

Hogan, J. W. and Laird, N. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in medicine*, **16**, 259–272.

Hoover, D., Graham, N., Chen, B., Taylor, J., Phair, J., Zhou, S., and Munoz, A. (1992). Effect of cd4+ cell count measurement variability on staging hiv-1 infection. *Journal of Acquired Immune Deficiency Syndrome*, **5**, 794–802.

Jacqmin-Gadda, H., Thiebaut, R., Chene, G., and Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, **1**, 355–368.

Jeong, J. and Fine, J. (2006). Direct parametric inference for the cumulative incidence function. *Applied Statistics*, **55**, (2),187–200.

Jewell, N. and Kalbfleisch, J. (1996). Marker processes in survival analysis. *Lifetime Data Analysis*, **2**, No.1, 15–29.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics.

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

Kiuchi, A., Hartigan, J., Holford, T., Rubinstein, P., and Stevens, C. (1995). Change points in the series of t4 counts prior to aids. *Biometrics*, **51**, 236–248.

Klein, J. and Andersen, P. (2005). Regression modeling of competing risks data based on pseudovales of the cumulative incidence function. *Biometrics*, **61**, 223–229.

Laird, N. (1982). Component of variance component using the em algorithm. *Journal of Statistical Computation and Simulation*, **14**, 295–303.

Laird, N. and Ware, J. (1982). Random effects in longitudinal data. *Biometrics*, **38**, 963–974.

Lange, N., Carlin, B., and Gelfand, A. (1992). Hierarchical bayes models for the progression of hiv infection using longitudinal cd4 t-cell numbers. *Journal of the American Statistical Association*, **87**, 615–632.

Larson, M. and Dinse, G. (1985a). A mixture model for the regression analysis of competing risks data. *Applied Statistics*, **34**, 201–211.

Larson, M. and Dinse, G. (1985b). A mixture model for the regression analysis of competing risks data. *Applied Statistics*, **34**, 201–211.

Ledergerber, B., Egger, M., and Erard, V. (1999a). Aids-related opportunistic illnesses occurring after initiation of potent antiretroviral therapy. *Journal of the American Medical Association*, **282**, 2220–2226.

Ledergerber, B., Egger, M., and Opravil, M. (1999b). Clinical progression and virological failure on highly active antiretroviral therapy in hiv-1 patients: a prospective cohort study. *Lancet*, **353**, 863–868.

Lepri, A., Phillips, A., Monforte, A., Castelli, F., Antinori, A., deLuca, A., Pezzotti, P., Alberici, F., Cargnel, A., Grima, P., Piscopo, R., Prestileo, T., Scalise, G., Vigevani, M., Moroni, M., and Group, I. S. (2001). When to start highly active antiretroviral therapy in chronically hiv-infected patients: evidence from icona study. *AIDS*, **15**, 983–990.

Liang, H. and Zou, G. (2007). Analysis of relation between virologic responses and immunologic responses, patient's factors in aids clinical trials using a semiparametric mixed-effects model. *Biometrical Journal*, **49**, 3, 406–415.

Liang, H., Wu, H., and Carroll, R. (2003). The relationship between virologic and immunologic responses in aids clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, **4**, 297–312.

Lindley, D. and Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–42.

Lunn, M. and McNeil, D. (1995). Applying cox regression to competing risks. *Biometrics*, **51**, 524–532.

Marubini, E. and Valsecchi, M. (2004). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley: New York.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Model*. Chapman & Hall Ltd.

Miller, V., Staszewski, S., Nisius, G., Lepri, A., and Phillips, C. S. A. (1999). Risk of new aids diseases in people on triple therapy. *Lancet*, **353**, 463–464.

Molla, A., Korneveva, M., Gao, Q., Vasavanonda, S., Schipper, P., Mo, H., Markowitz, M., Chernyavskiy, T., Niu, P., Lyons, N., Hsu, A., Granneman, G., Ho, D., Boucher, C., Leonard, J., Norbeck, D., and Kempf, D. (1996). Ordered accumulation of mutations in hiv protease confers resistance to ritonavir. *Nature Medicine*, **2**, 760–766.

Ng, S. and McLachlan, G. (2003). An em-based semi-parametric mixture model approach to the regression analysis of competing risks data. *Statistics in Medicine*, **22**, 1097–1111.

Opravil, M., Lederberger, B., Furrer, H., Hirschel, B., Imhof, A., Gallant, S., Wagels, T., Bernasconi, E., Meienberg, F., Rickenbach, M., Weber, R., and the Swiss HIV Cohort Study (2002). Clinical efficacy of early initiation of haart in patients with asymptomatic hiv infection and cd4 cell count $>$ 350x1000000/l. *AIDS*, **16**, 1371–1381.

Pantazis, N. and Touloumi, G. (2005). Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Applied statistics*, **54**, Part2, 405–423.

Pepe, M. and Mori, M. (1993). Kaplan meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, **12**, 737–751.

Phillips, A., Staszewski, S., Weber, R., Kirk, O., Francioli, P., Miller, V., Vernazza, P., Lundgren, J., and Ledergerber, B. (2001). Hiv viral load response to antiretroviral therapy according to the baseline cd4 cell count and viral load. *Journal of the American Medical Association*, **286**, 2560–2567.

Pomerantz, R. (1995). Time to hit hiv, early and hard. *New England Journal of Medicine*, **333**, 450–451.

Pomerantz, R. (2001). Initiating antiretroviral therapy during hiv infection: confusion and clarity. *Journal of the American Medical Association*, **286**, 2597–2599.

Potthoff, R. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

Press, W., Teutolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in FORTRAN. The Art of Scientific Computing (2nd edn)*. Cambridge University Press: New York.

Putter, H., Fiocco, M., and Geskus, R. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in medicine*, **26**, (11), 2389–2430.

Rao, C. (1965). The theory of least square when the parameters are sthocastic and its application to the analysis to growth curves. *Biometrika*, **52**, 447–458.

Shi, M., Weiss, R., and Taylor, J. (1995). An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied statistics*, **45**, 151–163.

Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modelling of longitudinal and time-to-event data. *Biometrics*, **58**, 742–753.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measure of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 3, 583–639.

Sy, J., Taylor, J., and Cumberland, W. (1997). A stochastic model for the analysis of bivariate longitudinal aids data. *Biometrics*, **53**, 542–555.

Tai, B., Machin, D., White, I., and Gebski, V. (2001). Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. *Statistics in medicine*, **20**, 661–684.

Taylor, J., Cumberland, W., and Sy, J. (1994). A stochastic model for longitudinal aids data. *Journal of the American Statistical Association*, **89**, 727–736.

Thiebaut, R., Jacqmin-Gadda, H., Babiker, A., Commenges, D., and collaboration, T. C. (2005). Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of cd4+ cell count and hiv rna viral load in response to treatment of hiv infection. *Statistics in medicine*, **24**, 65–82.

Thiebaut, R., Jacqmin-Gadda, H., Walker, S., Sabin, C., Prins, M., Amo, J. D., Porter, K., Dabis, F., Chene, G., and Collaboration, T. C. (2006). Determinants of response to first haart regimen in antiretroviral-naive patients with an estimated time since hiv seroconversion. *HIV Medicine*, **7**, 1–9.

Tjalling, J. (1995). Historical development of the newton-raphson method. *SIAM Review*, **37**, (4), 531–551.

Touloumi, G., Pocock, S., Babiker, A., and Darbyshire, J. (1999). Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in medicine*, **18**, 1215–1233.

Touloumi, G., Pocock, S., Babiker, A., and Darbyshire, J. (2002). Impact of missing data due to selective dropouts in cohort studies and clinical trials. *Epidemiology*, **13**, 347–355.

Tsiatis, A., DeGruttola, V., and Wulfsohn, M. (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.

Wang, Y. and Taylor, M. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of American Statistical Association*, **455**, 895–905.

Ware, J. (1983). Growth curves. *Encyclopedia of Statistical Science, Kotz,s., Johnson, N.L., and read, C.B.*

Wu, M. and Carroll, R. (1988). Estimation and comparison of changes in presence of onformative right censoring by modelling the censoring process. *Biometrics*, **44**, 175–188.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, No.1, 330–339.

Zeger, S. and Diggle, P. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, **50**, 689–699.

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

Zeger, S. and Liang, K. (1986). Longitudinal data analysis for discrete and continuos outcomes. *Biometrics*, **42**, 121–130.

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

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795–802.