**MASTER OF BIOSTATISTICS**

**NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS**
MEDICAL SCHOOL
DEPARTMENT OF MATHEMATICS

**UNIVERSITY OF IOANNINA**
DEPARTMENT OF MATHEMATICS

THESIS
ALKIVIADIS OIKONOMIDIS

# THE USE OF DIFFERENT TIMESCALES IN MULTIPLE EVENT SURVIVAL ANALYSIS WITH APPLICATION TO CARE ADHERENCE IN HIV-POSITIVE PATIENTS OF THE GREEK COHORT STUDY AMACS

ATHENS, 2017

This thesis is submitted in partial fulfilment of the requirements for the degree of

MASTER OF BIOSTATISTICS

that is given by the Medical School and the Department of Mathematics of the National and Kapodistrian University of Athens and the Department of Mathematics of the University of Ioannina.

It was approved on _____________ by the thesis examining committee:

| | | |
|---|---|---|
| G. Touloumi (supervisor) | Professor | _____________ |
| F. Siannis | Assistant Professor | _____________ |
| V. A. Sypsa | Assistant Professor | _____________ |

## Acknowledgements

I would like to express my sincere gratitude to all people who supported and helped me through the process of this thesis.

## Abbreviations and acronyms

| | |
|---|---|
| AIDS | Acquired Immunodeficiency syndrome |
| AMACS | Athens Multicenter AIDS Cohort Study |
| CASCADE | Concerted Action on SeroConversion to AIDS and Death in Europe |
| c.d.f. | cumulative distribution function |
| COHERE | Collaboration of Observational HIV Epidemiological Research Europe |
| df | degrees of freedom |
| ECDC | European Centre for Disease Prevention and Control |
| i.i.d. | independent and identically distributed |
| HAART | Hyper Active Antiretroviral Therapy |
| HCDCP | Hellenic Centre for Disease Control and Prevention |
| MLE | Maximum Likelihood Estimation |
| MPnLE | Maximum Penalized Likelihood Estimation |
| NGO | Non-governmental organization |
| p.d.f. | probability density function |
| PVF | Power Variance Function |
| PWID | People Who Inject Drugs |
| WHO | World Health Organisation |

## Definitions

| | |
|---|---|
| $\beta$ | the fixed effects vector |
| $\Gamma( . )$ | Gamma function |
| $\delta_{ij}$ | Failure indicator |
| $f( . )$ | Density function |
| $F( . )$ | Cumulative density function |
| $g( . )$ | Density function for censoring times |
| $G( . )$ | Cumulative density function for censoring times |
| $\lambda( . )$ | Hazard function |
| $\Lambda( . )$ | Cumulative hazard function |
| $L( . )$ | Likelihood function |
| $t_{ij}$ | Event times |
| $S( . )$ | Survival function |
| $Z_{ij}$ | Covariate vector |
| $x_{ij}$ | Survival times |
| $u_i$ | Frailty term |

# Table of Contents

# Index of Figures

# Index of Tables

# Chapter 1.

# Introduction

## 1.1. HIV

The Human Immunodeficiency Virus (HIV) is a lentivirus that belongs to the subgroup of retroviruses and targets the immune system. Being a retrovirus means that in order to replicate, HIV needs to synthesize a DNA copy of its RNA. Acquired Immunodeficiency Syndrome (AIDS) is the most progressed stage of the HIV infection with a latency period of 2 to 15 years in the absence of antiretroviral therapy.

HIV infects vital cells of the human immune system, which are called CD4+ cells and are a type of T lymphocyte cell (T cell), causing its progressive impairment. The infected organism then becomes vulnerable to opportunistic infections and even some types of cancer that eventually lead to death since the weakened immune system is unable to fight off these threats.

HIV can be found in human bodily fluids including blood, semen, vaginal and anal fluids as well as breast milk; hence it can be transmitted in several ways through the exchange of these fluids. Having unprotected sexual intercourse may be the most common way to become infected, however HIV can be also transmitted via sharing contaminated syringes, receiving contaminated blood transfusions or between a mother and her child during pregnancy, labor and breast feeding. Contrary to common beliefs in the previous years, HIV cannot be transmitted through saliva, sweat or urine. (AVERT, 2015)

The course of infection, described in a relatively simple way, is that once the virus enters the body, seeks the CD4+ cells and attaches itself to their surface. The viral envelope then fuses with the cell membrane and the genome of the virus is released into the cell. Once inside the cell, the genetic material of HIV, which is called HIV RNA, is transcribed into HIV DNA by the use of an enzyme called transcriptase, making the HIV DNA able to combine with the DNA of the cell. This HIV stage is called provirus. Provirus begins the production of new viruses by using the host cell's reproduction mechanism, which will later infect other cells (NIH, 2015).

By the end of 2014, the estimated number of people living with HIV globally was 36.9 [34.3 – 41.4] million, with approximately 2 [1.9 – 2.2] million of those people being newly infected patients. In the same year, the accumulated number of deaths credited to HIV and HIV related causes reached 34 million.

In Eastern Europe and Central Asia there were 140000 HIV infections diagnosed within 2014, which consist a 30% increase between 2000 and 2014. According to the European Centre for Disease Prevention and Control (ECDC) this is the highest number of new cases reported in one year since mandatory case reporting was implemented in 1980. It is indicated by the WHO Regional Office for Europe that the outbreak of HIV is driven by the growing trend in the eastern part of Europe where, the number of newly reported cases has doubled.

According to the Joint United Nations Programme on HIV/AIDS (UNAIDS), there were 2.3 million people living with HIV in Western & Central Europe and North America by the end of 2014. In the same year, 88000 new infection were reported, of them 29992 were reported by the 31 EU/EEA countries. Sex between men remains the principal means of transmission, accounting for 42% of all HIV infections that were diagnosed in 2014.

Moreover, in certain EU countries (Bulgaria, Czech Republic, Hungary and Malta) there is a more than 200% increase in the rates of new diagnoses since 2005 whereas in other members of the EU, rates present a downward trend of more than 25% (Austria, Estonia, the Netherlands, France).

## 1.2. HIV/AIDS in Greece

In Greece, Infectious Disease Units, AIDS Reference Centres and Hospitals must report any new HIV infection to the Hellenic Center for Disease Control and Prevention (HCDCP) which among other things is in charge of issuing annual reports about the characteristics of the epidemic. The epidemic of HIV in Greece, like many other countries in the European Union, is considered stable, low-level and concentrated in key populations, particularly in men who have sex with men (MSM). However, in the recent years, along with the financial crisis that the country is experiencing since 2008, an on-going HIV outbreak has been observed. Although the extent to which the crisis has affected the HIV outbreak is yet uncertain, the numbers and facts are indicative.

In 2012 for the first time in Greece, the incidence of HIV infected individuals among people who inject drugs (PWID) exceeded the new cases of HIV infections reported among MSM. In the span of only two years, a huge increase in the number of reported new cases among PWID was observed, as in 2010 there were 16 new reports in 2011 there were 266 and finally in 2012 the number was as high as 551.

Although there is a combination of several factors that explain why this outbreak took place among PWID, the lack of preventive services seems to be the most significant. Interventions like opioid treatment and needle/syringe programs, that were implemented by the Greek authorities and NGOs in order to target the restraint of the outbreak seem to have had effect. There has been a gradual reduction in the amount of new reported cases among PWID from 2012 (n=522) to 2013 (n=262) and 2014 (n=106), while this trend seems to hold also during the first three quarters of 2015 (compared to those from the previous years). This shows how important these interventions are and how the withdrawal of HIV prevention program is related to the increase of HIV prevalence.

Nevertheless, the main route of HIV transmission in Greece remains the male to male sex with a total of 7003 (46,34%) reported cases between January 1, 1981 and October 31,2015. Men are seen to be more affected by the epidemic with 82,41% of the total cases being credited to them. In addition the most affected age group is the one between 30-35 years old, having 20,5% of all the reported cases by the end of 2014.

As far as AIDS is concerned, case reporting was implemented in 1984 and by the end of 2014 there were 3661 reported cases, meaning that up to this point, 25.36% of the total 14434 HIV-positive individuals had developed AIDS. Of these, 3042 (84,2%) were males and the remaining 579 (15.8%) were females. By the end of 2014, when the most recent accumulated data are available, the age group between 30 and 49 was the predominant group in developing AIDS.

The need of reporting both AIDS and HIV cases, which is a key factor in the epidemiological surveillance, arose especially after the implementation of antiretroviral therapies (ART) because the use of ART delayed the onset of AIDS. As far as Greece is concerned, HIV reporting started in 1998, 14 years after AIDS started being reported.

Finally, since 1983 when the first death credited to AIDS was reported, there have been reported 1862 deaths due to AIDS or AIDS related causes in total. Of these cases 1624 (87,2%) were males and 238 (12,8) females.

## 1.3. Hyper Active Antiretroviral Treatment

There might not be a cure for HIV yet, however there are effective ways of controlling the virus thus allowing people living with it to enjoy a longer and healthier life. Hyper Active Antiretroviral Treatment (HAART) consists of a combination of three or more antiretroviral (ARV) drugs (WHO, 2015b)· Since 1995 when HAART was first introduced to the public, it has been proven to be the most effective way to suppress the virus load.

The significance of HAART is reflected on the gradual increase of the ART eligibility criteria based on the WHO recommendations over the past years; patients now become eligible at higher CD4+ cell counts than before. In order to achieve the Fast-track goals of limiting the new HIV infections down to 200000 by the year 2030; goals that have been set by the Joint United Nations Programme on HIV/AIDS (UNAIDS) and will eventually lead the epidemic to an end (UNAIDS (2014)). Since 2002, when WHO published its first guidelines on when to start ART, the eligibility limits have been increased from 200 CD4 cells per $mm^3$ (WHO, 2002) to 500 CD4 cells per $mm^3$ in 2013 (in adults and adolescents regardless their HIV clinical stage). On September of 2015, an early release of revised guidelines were publicized setting every human living with HIV eligible for initiating ART regardless their CD4 cell count or their WHO clinical stage.

In addition, HAART can also be beneficial when used as part of a prevention strategy. In the study trial performed by Cohen M.S. et al (2011) it has been shown that a nearly 96% relative reduction in the number of linked HIV-1 transmissions between serodiscordant couples can be succeeded by the use of early antiretroviral therapy. A serodiscordant couple is when two people are in a continuing sexual relationship and one of them is HIV positive while the other is not. As a result, is was recommended by WHO in 2013, that all HIV positive partners in serodiscordant couples should be offered ART regardless of their CD4 cell count (WHO, 2013).

Furthermore, HAART is also used as part of the strategy for preventing mother-to-child HIV transmission (PMTCT) since vertical or mother-to-child transmission (MTCT) is the main source of child infection. In the 2013 WHO guidelines it was recommended that all child bearing and breastfeeding HIV infected women should initiate ART regardless of their clinical eligibility criteria, while in the 2015 WHO it was added that ART provision should be carried on even after the cessation of breastfeeding.

According to WHO, as of March 2015, among the approximately 37 million people who live with HIV, 15.8 million people were receiving HAART, representing almost 42% of those in need. The corresponding estimated percentage in the Western/Central Europe & North America is 51% of all people living with HIV.

## 1.4. AMACS

The Athens Multicenter AIDS Cohort study (AMACS) is an ongoing population cohort study initiated in 1996 (the following year after HAART availability) in order to establish a large database of HIV-1 positive individuals in Greece.

The main objectives of the AMACS are to evaluate the long-term efficiency of HAART, to describe any potential trends in the prevalence or frequency of AIDS defining events and related deaths as well as to locate any association between risk factors and virologic or immunologic response to HAART.

The AMACS is comprised by the 11 largest HIV-1 clinics based in Athens along with two recently added clinics, one in southern and one in northern Greece, Rio and Alexandroupolis respectively. According to the data of the Hellenic Centre for Disease Control and Prevention, between January 1, 1984 and October 31, 2015, 15109 individuals have been reported to be infected by HIV. Out of these people, 7849 are or have been monitored by the clinics that participate in the AMACS.

The criteria for inclusion in the AMACS of an HIV positive individual that is monitored in the participating clinics were to be alive on January 1, 1996 and also to be monitored for at least one year in the same clinic. Apart from the socio demographic characteristics and medical history of the participants that is collected upon the enrolment in the cohort, additional prospective data are collected during the following visits in the clinics. Such data can be the CD4 and CD8 count, viral load, antiretroviral therapy (including reasons for change or serious adverse events), other relative infections that might occur, the clinical stage of AIDS, HIV seroconversion, results from laboratory tests (biochemical, blood, urinal) and many other. The AMACS study has been approved by the Athens University IRB, the HCIDC IRB, the National Organization for Medicines and the National Ethics Committee. In accordance to the study ethics and data protection policy, data are provided by the conducting clinics to AMACS anonymously.

Finally, AMACS participates in the Collaboration of Observational HIV Epidemiological Research Europe (COHERE) as well as also in the Concerted Action on SeroConversion to AIDS and Death in Europe (CASCADE). COHERE is a collaboration that was formed in order to harmonize existing longitudinal data on HIV-positive persons collected across Europe to answer key research questions that could not be addressed adequately by individual cohorts. CASCADE is also a collaboration between the investigators of 29 cohorts of patients with well-estimated dates of HIV seroconversion across Europe, Australia, Canada, and Africa.

# Chapter 2.

# Survival Analysis

In this introductory chapter we try to clarify some of the basic concepts of survival analysis and the notation that we are going to use as basis for the presentation of the random-effect models, that we are going to call frailty models from now on.

## 2.1. Survival Data

Survival data is another way to refer to time-to-event data, where we measure the time until an event occurs. As Kalbfleisch and Prentice (2002) explain, it is of high importance to have a strict definition of time origin and what constitutes an event or end-point. In many cases, death is considered as the event of interest hence the term Survival analysis. Survival analysis can be applied in many disciplines such as economics, engineering, sociology, biology and medicine. In this thesis we focus on the application of survival analysis methods to medicine.

In the medical setting, the term survival analysis normally refers to the modelling of time to death (since death is usually regarded as the event). However, there are many occasions where an event can be considered to be the occurrence of a disease or the recurrence of a symptom, such as epileptic seizures; in other words, event is the transition from a state to another (Hougaard, 2000). In the example of death, the transition is from the state of being alive to the state of being dead and in the example of the occurrence of a disease, the transition is from the state of being healthy to the state of being unhealthy.

But why do we have to develop special techniques for assessing these types of data? The main reason is that we observe something that develops dynamically over time, the survival time. Additionally, the event of interest might not occur within the observational interval and so the information that we have is limited and is called censored. Therefore, standard statistical approaches are not suitable and we need a special statistical theory. For the observational time, we consider a non-negative random variable T that represents the period from a well defined time origin until the occurrence of the event of interest or until censorship.

## 2.2. Basic notation

### 2.2.1. Density and Cumulative Distribution function

As we mentioned before, let T be a non-negative random variable from a continuous distribution. Then we denote the probability density function (p.d.f.) which defines the distribution of T uniquely and completely by f :

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leqslant T \leqslant t + \Delta t) \tag{2.2.1}$$

and the corresponding cumulative distribution function (c.d.f.) by F:

$$F(t) = P(T \leqslant t) = \int_0^t f(u)\,du \tag{2.2.2}$$

### 2.2.2. Survival function

In survival analysis the survival function S(t) is of more interest than the cumulative distribution function and that is because the main concern here is to calculate the probability of an observational unit “surviving” beyond time t. So we define the survival function as

$$S(t)=1-F(t)=P(T>t)=\int_{t}^{\infty} f(u)\,du \tag{2.2.3}$$

And so we get :

$$f(t)=-\frac{\partial S(t)}{\partial t} \tag{2.2.4}$$

### 2.2.3. Hazard function

Another quantity in which one is more interested when discussing the survival analysis setting is the hazard function. Depending on the discipline that it is applied to, the hazard function gets different names such as mortality rate in demography, failure or incidence rate in epidemiology or the inverse of Mill's ratio in economics. In our case we will refer to the hazard function as the instantaneous failure rate or simply hazard rate. The hazard rate is defined by the conditional probability of failure within the interval (t, t + Δt] given the fact that the event has not occurred yet. The expression through which we obtain the rate is the limit of this conditional probability divided by the time interval Δt, where Δt tends to zero (Duchateau, 2008).

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{Pr(t\leqslant T\leqslant t+\Delta t\vee T\geqslant t)}{\Delta t}=\frac{f(t)}{S(t)} \tag{2.2.5}$$

At this point we can also define the cumulative hazard function Λ(t) which is related to the hazard rate and it is used to derive:

$$\Lambda(t)=\int_0^t \lambda(u)\,du \qquad (2.2.6)$$

from (2.5) we derive:

$$\lambda(t)=\frac{f(t)}{S(t)}=-\frac{\partial \log[S(t)]}{\partial t} \qquad (2.2.7)$$

and so expression (2.6) by using that S(0)=1, becomes:

$$\Lambda(t)=-\int_0^t \frac{\partial \log[S(u)]}{\partial u}du=-\log[S(t)]+\log[S(0)]=-\log[S(t)] \qquad (2.2.8)$$

finally we get another useful expression:

$$S(t)=e^{-\Lambda(t)}=e^{-\left(\int_0^t \lambda(u)du\right)} \qquad (2.2.9)$$

## 2.3. Censoring and Truncation

Survival data are often censored, truncated or even both in certain cases and as we mentioned before, that creates some implications in handling such data. Censoring and truncation give partial information about the observational time (Hosmer & Lemeshow, 2008), therefore, typical statistical methodology is not appropriate for this type of data; in the presence of censoring and/or truncation the form of the likelihood function becomes more complex. Thereby, we need to take into consideration the nature of our data in order to avoid mistakes in the form of the likelihood that we are going to use as a foundation for our inference.

*Censoring* comes in three schemes: left, right and interval which is a combination of right and left censoring. *Left censoring* arises when the only information we have is that the event of interest occurred at some point prior to the start of study, *right censoring* when the event of interest has not occurred by the end of the study period and finally, *interval censoring*, when all is known is that the event occurred within some interval.

We will focus on the right censoring scheme since it is the one that is present in almost every case that survival data is involved. The most common reason for partial information about an observation is that the observational unit has not failed until the end of the follow up period. For instance, in a longitudinal study of coronary heart disease, if a given participant of the study population has not developed the disease at the end of the study then his survival time is considered censored.

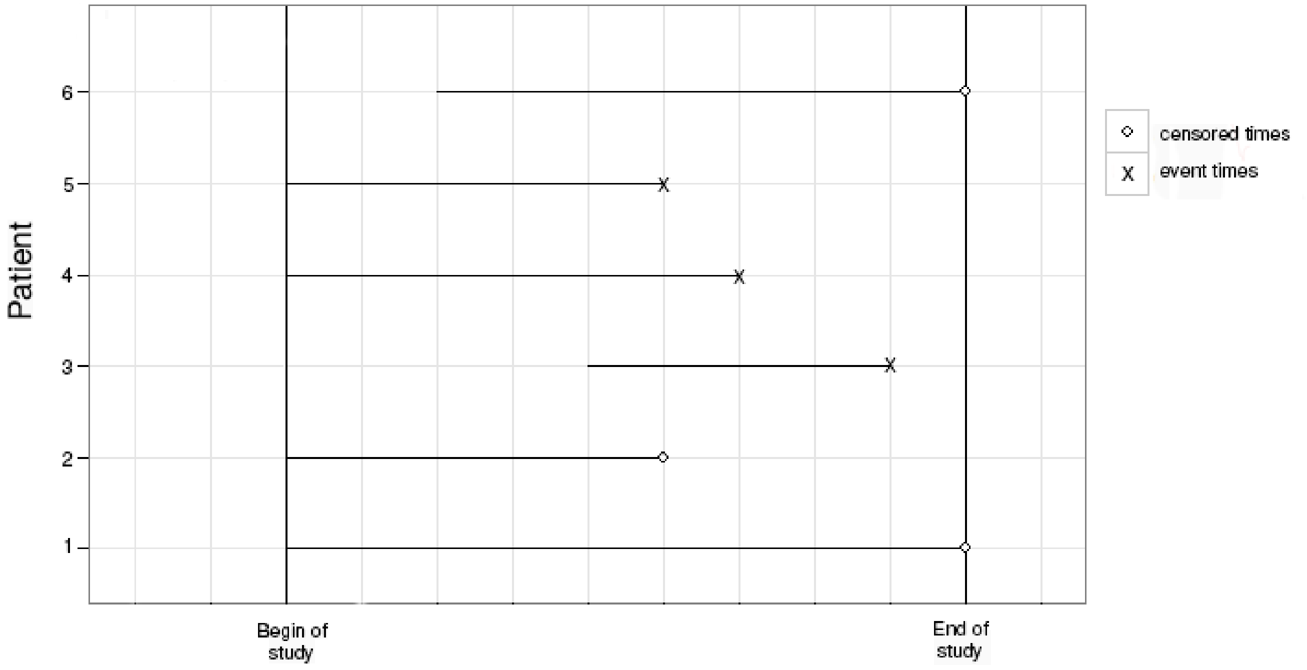
Apart from the termination of the study there are also other reasons where an observation is considered censored (Kleinbaum & Klein, 2006) as in the case of right censoring. It generally occurs due to three main reasons:

- Withdrawal for the study. A person can withdraw from the study for many reasons, such as adverse side effects from the treatment or not receiving satisfying results from treatment.
- Loss to follow-up. The researchers may lose contact with some of the participants of a study and then these people are considered lost to follow up and their survival times censored.
- Competing risks. We are not able to observe the failure of an individual since another event occurred before. In our example of the coronary heart disease study, where heart failure is the event of interest, if one dies by traffic accident, which is not related to the outcome of interest, then his survival time is considered censored.

At this point, we need to mention that the cause of censoring should not be related to the event of interest otherwise this can introduce bias into the estimation of survival times. However, it is difficult to be certain whether censoring and the event of interest are independent or not. Likewise, we often need to assume this independence, meaning that censoring is assumed to be independent when the failure rates that we observe in the presence of censoring would be the same as if there wasn't any censoring (Kalbfleisch & Prentice, 2002). For instance, in the last example with the traffic accident, the person involved might have had a heart incident that led to the accident. Therefore, great attention should be given to the assumption to avoid introducing bias to our inferences.

In the figure 2.1 that follows, the survival time of 6 patients of a longitudinal study are illustrated. The time scale that is used is the calendar time that represents the actual moment that every patient joined the study. Patients 3, 4 & 5 experienced the event of interest during the study with patients 4 & 5 being part of the study from the beginning while patient 3 entered the study later. Patients 1, 2 & 6 have censored survival times, with patients 1 and 6 being still alive at the end of the study while patient 2 was lost to follow-up.

**Figure 2.1:** *Example of event and censoring times in a hypothetical longitudinal study (Calendar time)*



There is also another way to represent censoring and event times by rescaling the time to the time values that participants spent on the study as seen in Figure 2.2.

In paragraph 2.1 we denoted with T the non-negative random variable that represents the observational time. Now, let $X_1,X_2,...,X_n$ be independent and identically distributed (i.i.d.) survival times and $C_1,C_2,...,C_n$ be i.i.d. censoring times of n individuals under observation with cumulative distribution function F and G respectively, which are also continuous. If $T_i = \min\{ X_i , C_i \}$, then in the case of right censored data, we can only observe $(T_1,\Delta_1),(T_2,\Delta_2),...,(T_n,\Delta_n)$ where $\Delta_i$ is the failure indicator for subject i, i=1,...,n. For the failure indicator we have :

$$\Delta_i=\begin{cases} 1 \text{ if } X_i \le C_i & , T_i \text{ is not Censored} \\ 0 \text{ if } X_i > C_i & , T_i \text{ is Right Censored} \end{cases}$$

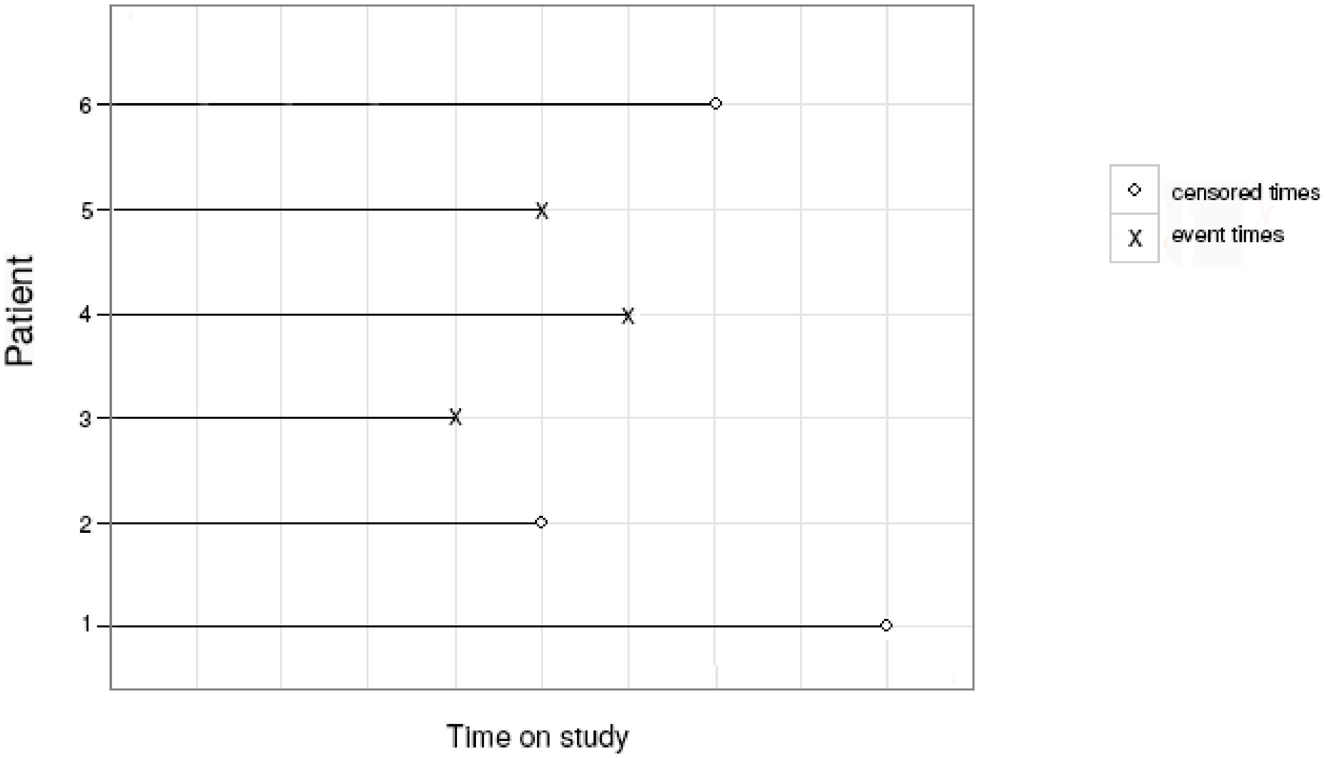


**Figure 2.2:** Rescaling the event and survival times of the figure 2.1 (Time on study)

As mentioned above, a major assumption that we need to make throughout this thesis is that the survival times $X_1,X_2,...,X_n$ are independent from the censoring times $C_1,C_2,...,C_n$. Therefore, if we denote by H the distribution function of the observational time $T_i$ following the notation by Wienke (2010), we have:

$$\begin{aligned} H(t) &= P(min\{T^*, C\} \leqslant t) \\ &= 1 - P(min\{T^*, C\} > t) \\ &= 1 - P(T^* > t, C > t) \end{aligned}$$

If now, we assume independence between censoring and event times, we have:

$$\begin{aligned} H(t) &= 1 - P(T^* > t)P(C > t) \\ &= 1 - (1 - P(T^* \leqslant t))(1 - P(C \leqslant t)) \\ &= 1 - (1 - F(t))(1 - G(t)) \end{aligned}$$

It should also be mentioned, that this is the so called Type I right censoring mechanism; furthermore, Type II right censoring occurs when the study terminates after a specific and predetermined number of failures k, where k<n and is mostly encountered in the engineering field, in order to check equipment's life.

Finally, *truncation,* which is commonly confused with censoring*,* comes in two schemes: left and right. We will mainly focus on left truncation which is encountered more frequently in the medical field. Generally, truncated data arise when we only observe event times which lie in an interval ($Y_L$ , $Y_R$). This must not be confused with the interval ($L_i$ , $R_i$) of interval censoring since in that case there is partial information in contrast to truncation where no information outside the interval is available.

*Right truncation* occurs when the entire population of the study has experienced the event of interest. Klein and Moeschberger (2003) use the example of an AIDS related study to illustrate that. The sample included only patients that had already developed AIDS after being infected by contaminated blood transfusion, when the registry was sampled, thus patients that had not yet developed AIDS where excluded from the study.

*Left truncation* occurs when the individuals under observation are already in risk before entering the study, subsequently those who have already experienced the event of interest are not observed. Therefore, under truncation, the distribution of the survival times is conditional to $T > \tau^*$, with $\tau^*$ being the non-random truncation time. Additionally, in left truncation, the upper limit of the interval is infinite or equal to the censoring time Δ.

## 2.4. Constructing the likelihood

While constructing the likelihood we always have to keep in mind that the form of the expression should represent all the information that is observed (Hougaard, 2000). We will show the way to construct the likelihood expression in the presence of right censored data with random censoring. As we mentioned before, the form of the likelihood is determined by the data that we have and in that case we have two possible outcomes, first that the observation unit fails and second that it is censored. We will see these two cases separately and then we will combine them for the final form of the likelihood (Duchateau & Janssen, 2008).

- An event time ($y_i = t_i$ , $\delta_i$=1) contributes to the likelihood:

$$\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}P\left(y_i-\varepsilon<Y_\iota<y_i+\varepsilon,\delta_i=1\right)$$

$$= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P\left(y_i - \varepsilon < Y_\iota < y_i + \varepsilon , T_i \leqslant C_i\right)$$

$$= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{y_i - \varepsilon}^{y_i + \varepsilon} \int_{t}^{\infty} dG\left(c\right) dF\left(t\right) \qquad \text{( due to independence )}$$

$$= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{y_i - \varepsilon}^{y_i + \varepsilon} \left(1 - G\left(t\right)\right) dF\left(t\right)$$

$$= \left(1 - G\left(y_i\right)\right) f\left(y_i\right) \tag{2.4.1}$$

- A right censored observation ($y_i = c_i$ , $\delta_i$=0) contributes to the likelihood:

$$\lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P\left(y_i - \varepsilon < Y_\iota < y_i + \varepsilon , \delta_i = 0\right)$$

$$= \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} P\left(y_i - \varepsilon < C_i < y_i + \varepsilon , T_i > C_i\right)$$

$$= \left(1 - F\left(y_i\right)\right) g\left(y_i\right) \tag{2.4.2}$$

By combining (2.4.1) and (2.4.2) we get the likelihood form:

$$L = \prod_{i=1}^{n} \left[\left(1 - G\left(y_i\right)\right) f\left(y_i\right)\right]^{\delta_i} \left[\left(1 - F\left(y_i\right)\right) g\left(y_i\right)\right]^{1 - \delta_i} \tag{2.4.3}$$

The expression of the likelihood (2.4.3) based on the assumption of non-informative censoring can be rewritten as:

$$L = \prod_{i=1}^{n} \left[f\left(y_i\right)\right]^{\delta_i} \left[S\left(y_i\right)\right]^{1 - \delta_i} \tag{2.4.4}$$

Non-informative censoring means, given the people at risk at a specific time, that the likelihood for the censored observations does not depend on the parameters of interest, thus it does not contain any information about them. So the factors $\left(1-G\left(y_i\right)\right)^{\delta_i}$ and $\left(g\left(y_i\right)\right)^{1-\delta_i}$ can be ruled out of the likelihood expression (Fleming & Harrington, 2011).

Another way to express (2.4.4) is by using the hazard function and more specifically the expressions (2.2.7) and (2.2.9) described before

$$\begin{aligned} L &= \prod_{i=1}^{n}\left[\lambda\left(y_i\right)S\left(y_i\right)\right]^{\delta_i}\left[S\left(y_i\right)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[\lambda\left(y_i\right)\right]^{\delta_i}S\left(y_i\right) \\ &= \prod_{i=1}^{n}\left[\lambda\left(y_i\right)\right]^{\delta_i}e^{-\left(\int_0^{y_i}\lambda(u)\,du\right)} \end{aligned} \tag{2.4.5}$$

## 2.5. Non parametric techniques in survival analysis

In this section, we focus on the simple non parametric techniques used to model the survival function. The “Non-parametric” term is used in order to emphasize that we do not need to make any assumption for the distribution of the survival times and we just rely on our data to make an inference. Of course this choice comes with a price, since non-parametric techniques usually require bigger samples to derive to reasonable inferences and the estimation of the hazard function cannot be carried out since the estimated distribution is discrete (Hougaard, 2000).

### 2.5.1. Kaplan Meier estimator

The Kaplan-Meier estimator is probably the most common approach to estimate the survival function when dealing with right censored data. It was proposed by Kaplan and

Meier (1958) and it is also called Product-Limit estimator.

Let's suppose that $y_1 < y_2 < \dots < y_k$ are the observed failure times of a n-sized sample from a homogeneous population. Let S also be the survivor function of the population. If $d_j$ individuals fail at time $y_j$ and $m_j$ individuals are right censored within the interval [ $y_j$ , $y_{j+1}$ ) at times $y_{j1},\cdots,y_{jm_j}$ , j=0,...,k, where $y_0 = 0$ and $y_{k+1}=\infty$. Additionally, let

$$n_j=\left(m_j+d_j\right)+\ldots+\left(m_k+d_k\right)=\sum_{i=j}^{k}\left(m_i+d_i\right)$$

denote the risk set of the sample population just before the j-th failure. Then the probability of a failure at the time $y_j$ is:

$$P\left(T=y_j\right)=S\left(y_j^-\right)-S\left(y_j\right)$$

so the contribution of an individual to the likelihood who failed at the time $y_j$ is assumed to be:

$$f\left(y_j\right)=-\left.\frac{\partial S\left(y\right)}{\partial y}\right|_{y=y_j}$$

while the contribution of an individual that is censored at time $y_{ij}$ is:

$$P\left(T>y_{jl}\right)=S\left(y_{jl}\right)$$

Accordingly, the likelihood of the data is:

$$L=\prod_{j=0}^{k}\left\{\left[S\left(y_j^-\right)-S\left(y_j\right)\right]^{d_j}\prod_{l=1}^{m_j}S\left(y_{jl}\right)\right\} \tag{2.5.1}$$

Then we denote with Ŝ the estimation for the survival function that maximizes the likelihood function and is called maximum likelihood estimation (MLE). The estimation of the survival function is discrete (Kalbfleisch & Prentice, 2002), with hazard components $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_k$ at $t_1$ , … , $t_k$ respectively. Consequently we can use for the survival function the expressions:

$$\widehat{S}\left(y_j\right)=\prod_{l=1}^{j}\left(1-\widehat{\lambda}_l\right) \tag{2.5.2}$$

$$\widehat{S}\left(y_j^-\right)=\prod_{l=1}^{j-1}\left(1-\widehat{\lambda}_l\right) \tag{2.5.3}$$

So if we want to express the likelihood as a function of λ=(λ$_1$, λ$_2$, ..., λ$_κ$), by using the expressions (2.5.2) and (2.5.3) we have:

$$L\left(\lambda\right) \quad =\prod_{j=1}^{k}\left\{\lambda_j^{d_j}\prod_{i=1}^{j-1}\left(1-\lambda_i\right)^{d_j}\prod_{i=1}^{j}\left(1-\lambda_i\right)^{m_j}\right\}$$

$$=\prod_{j=1}^{k}\left\{\frac{\lambda_j^{d_j}}{\left(1-\lambda_j\right)^{d_j}}\times\prod_{i=1}^{j}\left(1-\lambda_i\right)^{d_j+m_j}\right\}$$

$$=\prod_{j=1}^{k}\left\{\frac{\lambda_j^{d_j}}{\left(1-\lambda_j\right)^{d_j}}\right\}\times\prod_{j=1}^{k}\prod_{i=1}^{j}\left\{\left(1-\lambda_i\right)^{d_j+m_j}\right\}$$

$$=\prod_{j=1}^{k}\left\{\frac{\lambda_j^{d_j}}{\left(1-\lambda_j\right)^{d_j}}\right\}\times\left(1-\lambda_1\right)^{\left(d_1+m_1\right)+\ldots+\left(d_k+m_k\right)}\times\left(1-\lambda_2\right)^{\left(d_2+m_2\right)+\ldots+\left(d_k+m_k\right)}\times\left(1-\lambda_3\right)^{\left(d_3+m_3\right)+\ldots+\left(d_k+m_k\right)}\times\ldots$$

$$=\prod_{j=1}^{k}\left\{\frac{\lambda_j^{d_j}}{\left(1-\lambda_j\right)^{d_j}}\right\}\times\prod_{j=1}^{k}\left\{\left(1-\lambda_i\right)^{\sum_{i=j}^{k}\left(d_i+m_i\right)}\right\}$$

$$=\prod_{j=1}^{k}\left\{\lambda_j^{d_j}\left(1-\lambda_j\right)^{n_j-d_j}\right\}$$

It can be easily shown that the MLE of the λ$_j$ is $\widehat{\lambda_j}=d_j/n_j$ where (j = 1, … , k), so the estimator proposed by Kaplan and Meier is:

$$\hat{S}(t) = \prod_{j \vee y_j \leq t} \frac{n_j - d_j}{n_j} \tag{2.5.4}$$

The Kaplan-Meier estimator is a step function and it changes only when an event is observed, therefore if we have censored times bigger than or equal to the last event time the estimator does not reach 0. In cases like this the estimator cannot be defined.

## 2.6. Regression Models

Till now, we only mentioned non parametric techniques to estimate the survival function by using independent identically distributed data that imply a homogeneous population (Wienke, 2010). In this section we will present the way to model the hazard function, since in survival analysis one is more interested in the rate or hazard of failure at any given time after the start of the study period. When we want to compare two groups that are similar other than the treatment arm that they belong to as an exception, then nonparametric methods mentioned above can be used. But what happens when a number of explanatory variables and risk factors are recorded for each individual under study as applies in most practical applications that the population is not homogeneous? In such cases we need to implement other techniques such as regression models.

### 2.6.1. Proportional hazards model

The proportional hazards model is probably the most popular model used in survival analysis. Let λ(t|z) be the hazard function of an individual at time t and $z^T = (z_1 , z_2 , \dots , z_p )$ the covariate vector. The proportional hazards model assumes that:

$$\lambda\left(t|z\right)=\lambda_0\left(t\right)\psi\left(z,\beta\right) \tag{2.6.1}$$

where the $\lambda_0(t)$ is an arbitrary baseline hazard rate and $\psi(z,\beta)$ is a non-negative function that depends on the covariates. In most of the cases, the function that is used for $\psi(.)$ is the exponential $\psi(\zeta, \beta)=\exp(z^T\boldsymbol{\beta})$, with z being the vector of covariates and $\boldsymbol{\beta}$ being the corresponding vector of regression parameters which is defined as $\boldsymbol{\beta}=(\beta_1 , \beta_2 , ..., \beta_p)$ .

$$\lambda\left(t|z\right)=\lambda_0\left(t\right)\exp\left(z^T\beta\right) \tag{2.6.2}$$

As its name implies, the model assumes that the hazard between two groups is proportional. For instance, if we have two individuals that they differ in a specific element, let's say smoking habits, then we denote the hazards of the smoker $x_s$ and non-smoker $x_{NS}$ as $\lambda_S(t|z)$ and $\lambda_{NS}(t|z)$ respectively. In this case, there is a single dichotomous covariate Z in the model, where Z = 1 marks that one smokes and Z = 0 contrariwise. So the hazard ratio would be:

$$HR\left(t,x_S,x_{NS}\right)=\frac{\lambda_S\left(t|z\right)}{\lambda_{NS}\left(t|z\right)}=\frac{\lambda_0\left(t\right)\exp\left(\beta\right)}{\lambda_0\left(t\right)}=\exp\left(\beta\right) \tag{2.6.3}$$

In the final expression, time is not included, which implies that the hazard ratio does not change over time; this is the essence of the separation of the time and the observed covariates.

What is more, we observe that the baseline hazard $\lambda_0(t)$ is ruled out from the final expression. But one could wonder what does this term means. A simple answer to this question is that the baseline hazard is a term that describes the dependance of the hazard function on time t. Another answer would be that the baseline hazard is the hazard when all the other covariates have a value of 0.

There are two approaches to the baseline hazard in the proportional hazards model. The first one is the parametric approach where it is assumed that the baseline hazard follows a specific distribution. The second one is when the baseline hazard is left unspecified and in this case we call it semiparametric proportional hazards model.

### 2.6.2. Semiparametric proportional hazards model

This approach is called semiparametric since the covariates of the model have a parametric nature while there is not a parametric specification of the baseline hazard function. Its biggest advantage is the one implied in the previous section. The simplicity of the interpretation of the result in (2.6.2) which can be regarded as a relative risk ratio. The model was introduced by Cox in his seminal paper (1972) where he suggested a partial likelihood approach.

As we showed in expression (2.4.5) of the likelihood, the form contains the hazard function. Under the proportional hazards model, the likelihood contains the unspecified hazard function along with the covariate function. As a result, we can now rewrite the likelihood (2.4.5) as:

$$L=\prod_{i=1}^{n}\left[\lambda_0\left(t_i\right)\exp\left(\beta Z_i\right)\right]^{\delta_i}\exp-\left(\int_0^{t_i}\lambda_0\left(u\right)\exp\left(\beta Z_i\right)du\right)$$

$$L=\prod_{i=1}^{n}\left[\lambda_0\left(t_i\right)\exp\left(\beta Z_i\right)\right]^{\delta_i}\exp-\left(\Lambda_0\left(t_i\right)\exp\left(\beta Z_i\right)\right) \qquad (2.6.4)$$

However, this expression remains problematic since it contains the baseline hazard $\lambda_0(t)$. In order to explain the partial likelihood approach we need to introduce a new notation. Assuming that there are no ties, we have the ordered event times $y_{(1)} < y_{(2)} < \dots < y_{(r)}$ with r=d (meaning that r is the number of events) and corresponding covariates $z_{(1)}, \dots, z_{(r)}$. So now we rewrite once again the likelihood as:

$$L=\prod_{i=1}^{r}\left[\lambda_0\left(y_{(i)}\right)\exp\left(\beta z_{(i)}^t\right)\right]\times\prod_{j=1}^{n}\exp-\left(\Lambda_0\left(y_j\right)\exp\left(\beta z_j^{\,t}\right)\right) \qquad (2.6.5)$$

Additionally we assume the discrete version of cumulative baseline hazard since the baseline hazard function is equal to 0 apart from the times that we observe an event. Accordingly we have:

$$\Lambda_0^{DIS}\left(y_j\right)=\sum_{y_{(i)}\leqslant y_j}\lambda_0\left(y_{(i)}\right) \qquad (2.6.6)$$

Combining the two last expressions (2.22) and (2.23) we get the survival likelihood:

$$\begin{aligned}L\left(\lambda_{(1)},\dots,\lambda_{(r)}\vee\beta\right) &=\prod_{i=1}^{r}\left[\lambda_0\left(y_{(i)}\right)\exp\left(\beta z_{(i)}^t\right)\right]\times\prod_{j=1}^{n}\exp-\left(\Lambda_0\left(y_j\right)\exp\left(\beta z_j^{\,t}\right)\right)\\ &=\prod_{i=1}^{r}\left[\lambda_0\left(y_{(i)}\right)\exp\left(\beta z_{(i)}^t\right)\right]\times\prod_{j=1}^{n}\exp\left\{-\sum_{i\vee y_{(i)}\leqslant y_j}\lambda_0\left(y_{(i)}\right)\exp\left(\beta z_j^{\,t}\right)\right\}\\ &=\prod_{i=1}^{r}\left[\lambda_0\left(y_{(i)}\right)\exp\left(\beta z_{(i)}^t\right)\right]\times\exp\left\{-\lambda_0\left(y_{(1)}\right)\sum_{j\vee y_j\geqslant y_{(1)}}\exp\left(\beta z_j^{\,t}\right)-\dots-\lambda_0\left(y_{(r)}\right)\sum_{j\vee y_j\geqslant y_{(r)}}\exp\left(\beta z_j^{\,t}\right)\right\}\\ &=\prod_{i=1}^{r}\left[\lambda_0\left(y_{(i)}\right)\exp\left(\beta z_{(i)}^t\right)\right]\times\exp\left\{-\sum_{i=1}^{r}\lambda_0\left(y_{(i)}\right)\sum_{j\in R\left(y_{(i)}\right)}\exp\left(\beta z_j^{\,t}\right)\right\} \qquad (2.6.7)\end{aligned}$$

We will now take the partial derivatives of the likelihood (2.6.6) with respect to $\lambda_0(y_{(i)})$, i=1,...,r , set all the equations to 0 and then by solving for $\lambda_0(y_{(i)})$ we get:

$$\lambda_0\left(y\left(i\right)\right)=\frac{1}{\sum\exp\left(z_j^t\beta\right)},i=1,\dots,r \qquad (2.6.8)$$

Finally if we replace in the likelihood expression (2.6.6) the result we got above then we reach to the partial likelihood expression of Cox that doesn't contain any of the parameters or the factor $e^{-d}$

$$L\left(\beta\right)=\prod_{i=1}^{r}\frac{\exp\left(z_{(i)}^{t}\beta\right)}{\sum\exp\left(z_{j}^{t}\beta\right)} \tag{2.27}$$

The expression (2.6.8) is called partial likelihood and it is used to estimate β through the well known maximization process. The justification along with the properties of Cox's partial likelihood approach are well documented (Gill, 1984; Fleming & Harrington, 1991).

# Chapter 3.

# Frailty Models

Frailty models were developed in an attempt to account for the effect of unmeasured variables that might affect the hazard function. The origin of the term frailty derives from gerontology, but there is no broadly accepted definition (Rockwood, 1999). A definition that has been promoted is that of Fried et al. (2001) which focuses on the physical frailty and considers it as a clinical syndrome assessed by the presence of three indicators: unintentional weight loss, exhaustion and low physical activity. Vaupel et al. (1979) introduced the term in the context of demography. In this paper, the population is considered heterogeneous and the frailty term is used to represent the heterogeneity among the groups in the analysis of mortality rates.

The frailty models are random-effect models that present two types of variation. The source of the first type is the random variation on the individual level which is described by the hazard function, and the second, is the group variation which is described by the frailty term. Frailty Models can be used in both univariate and multivariate survival analysis.

## 3.1. Univariate Frailty Models

In this section we are going to focus on the analysis of univariate data by the use of frailty models. The term univariate refers to the fact that the survival times are assumed to be independent, as the work of Vaupel et al. The basic assumption of this approach is that the population is homogeneous up to a point, however there are a number of explanatory variables and risk factors that make people differ greatly. Additionally, we would like to emphasize that it is impossible to quantify and measure all the factors that may affect the survival of a subject in a study.

The univariate frailty models are used since they can be more flexible than a standard regression model and also to describe the overdispersion in relation to the typical random variation (Klein et al., 2013).

Extending the model described in the expression (2.6.3) by including the frailty term we get

$$\lambda\left(t\middle|u,z\right)=\lambda_0\left(t\right)u_i\exp\left(z_i^T\beta\right),i=1,2,\ldots,n \qquad (3.1.1)$$

where $z_i^T$=($z_{1i}$, $z_{2i}$, … , $z_{ni}$) is the vector of covariates and **β** the respective regression parameters of the i-th subject. In the univariate case, one can easily see that this is a generalization of the proportional hazards model

## 3.2. Multivariate Frailty Models

The term multivariate survival data implies that the assumption of independence between survival times is no longer valid (Hougaard, 2000) Multivariate data are encountered when individuals form groups by sharing a common feature thus their survival times are correlated in some way. These groups arise in situations where individuals are related, such as family members, matched pairs of twins like the study on adult Danish twins (Hougaard et al., 1992) where the twins share a common childhood environment, especially during the pre-birth period. In cases like these, frailty is used in order to model the genetic effect.

In the multivariate analysis, frailty models are used to account for the heterogeneity between groups of subjects. This between-groups variability is accountable for the difference presented in the risks of the groups. The between-groups variability, in its turn, appears as dependence between the members of the groups who share common risks, which is why frailty models are used in the multivariate analysis.

We introduce the frailty term for each group so as to induce the dependence between the units within a cluster. The models processing this type of data are also called shared frailty models, since all the units within a cluster share a common factor, the frailty. Frailty is assumed to be responsible for the dependence mentioned above and its value is constant over time and common for all members of the cluster. When treating data from recurrent events, one should interpret the term "*shared"* as shared over time since every cluster is constituted by a single individual.

The biggest advantage of the frailty term being shared among different subjects is that it models the dependence between the survival times. In addition given the frailty, the time variables are conditionally independent, meaning that the observed times are independent if the frailty is integrated out. For instance, in the example of the twins study that was mentioned before, when the common genes are accounted for then the survival times of the twins are assumed to be conditionally independent. Generally the genes are unknown so their effect has to be integrated out but if we have some partial knowledge about the common genes then we can include them in the model as fixed effects.

### 3.2.1. Gamma Frailty Model

The frailty is a random-effect factor and in that sense, we need to specify a distribution for it (Duchateau & Janssen, 2008). Many different distributions have been proposed including the Gamma distribution, the positive stable, the inverse Gaussian and the compound Poisson; all these distributions belong to the power variance function (PVF) family. Even though in most of the cases the standard assumption is that the frailty follows a gamma distribution, there is no actual biological evidence supporting this choice but rather a mathematical reason which makes it appealing to the conductors of the studies. The frailty factor which is apparent in the conditional likelihood can be integrated out of the expression and then by relying on the classical likelihood maximization techniques we are able to obtain the estimates for the parameters of interest from the marginal likelihood and evaluate the distribution of the survival times.

The model assumes that there is independence between the groups but in the same time dependence between the times of the members of every group. That implies that if the frailty term presents no variation then the survival times are independent.

Therefore, we assume that the frailties u$_\iota$ are i.i.d from a Gamma distribution with mean 1 and unknown variance

$$u_i \sim \Gamma\left(\frac{1}{\theta},\frac{1}{\theta}\right) \qquad \text{with} \qquad E\left(u_i\right)=1 \qquad \text{and} \qquad Var\left(u_i\right)=\theta$$

and thus the probability density function is:

$$f\left(u\right)=\frac{u^{\left(1/\theta-1\right)}\exp\{-u/\theta\}}{\Gamma\left(1/\theta\right)\theta^{1/\theta}} \tag{3.2.1}$$

Based on the hazard function (3.1.1) and on the same technique that we used to derive the likelihood expression (2.6.4), the conditional likelihood for the i-th cluster is given by

$$L_i\left(\xi,\beta\middle|u_i\right)=\prod_{j=1}^{n_i}\left[\lambda_0\left(t_{ij}\right)u_i\exp\left(\beta z_{ij}^{t}\right)\right]^{\delta_{ij}}\exp-\left(\Lambda_0\left(t_{ij}\right)u_i\exp\left(\beta z_{ij}^{t}\right)\right) \tag{3.2.2}$$

Where ξ contains the baseline hazard parameters.

The marginal likelihood L$_{\text{marg, I}}$(ζ) for the i-th cluster is given by

$$L_{marg,i}\left(\zeta\right)=\int_0^\infty\prod_{j=1}^{n_i}\left[\lambda_0\left(t_{ij}\right)u\exp\left(\beta z_{ij}^{t}\right)\right]^{\delta_{ij}}\exp-\left(\Lambda_0\left(t_{ij}\right)u\exp\left(\beta z_{ij}^{t}\right)\right)\frac{u^{1/\theta-1}}{\theta^{1/\theta}\Gamma\left(1/\theta\right)}\exp\left(-u/\theta\right)du \tag{3.2.3}$$

with ζ=(ξ,θ,β)

There is a closed form expression for this integral (Duchateau, 2008) and we can rewrite this marginal likelihood as

$$L_{marg,i}\left(\zeta\right)=\frac{\prod_{j=1}^{n_i}\left(\lambda_0\left(t_{ij}\right)\exp\left(\beta z_{ij}^{\;t}\right)\right)^{\delta_{ij}}}{w^{1/\theta+d_i}\theta^{1/\theta}\Gamma\left(1/\theta\right)}\int_0^{\infty}\left(wu\right)^{1/\theta+d_i-1}\exp\left(-wu\right)d\left(wu\right) \tag{3.2.4}$$

by using the expression

$$w=1/\theta+\sum_{j=1}^{n_i}\Lambda_0\left(t_{ij}\right)\exp\left(\beta z_{ij}^{\;t}\right)$$

Now, if we work out the integral, we get

$$L_{marg,i}\left(\zeta\right)=\frac{\Gamma\left(d_i+1/\theta\right)\prod_{j=1}^{n_i}\left(\lambda_0\left(t_{ij}\right)\exp\left(\beta z_{ij}^{\;t}\right)\right)^{\delta_{ij}}}{\left(1/\theta+\sum_{j=1}^{n_i}\Lambda_0\left(t_{ij}\right)\exp\left(\beta z_{ij}^{\;t}\right)\right)^{1/\theta+d_i}\theta^{1/\theta}\Gamma\left(1/\theta\right)} \tag{3.2.5}$$

where $d_i=\sum_{j=1}^{n_i}\delta_{ij}$ is the number of observed events in cluster i (i=1,...,n).

Klein (1992) showed how one gets the marginal loglikelihood $l_{marg}(\zeta)$ by taking the logarithm and summing over the s clusters. The expression is given by

$$l_{marg}\left(\zeta\right)=\sum_{i=1}^{s}\Bigg[d_i log\theta-log\Gamma\left(1/\theta\right)+log\Gamma\left(1/\theta+d_i\right)$$
$$-\left(1/\theta+d_i\right)\log\left(1+\theta\sum_{j=1}^{n_i}\Lambda_0\left(t_{ij}\right)\exp\left(\beta z_{ij}^{\;t}\right)\right)+\sum_{j=1}^{n_i}\delta_{ij}\left(\beta z_{ij}^{\;t}+\log\lambda_0\left(t_{ij}\right)\right)\Bigg] \tag{3.2.6}$$

By maximising this log-likelihood, maximum likelihood estimators for ξ, θ and β can be derived.

### 3.2.2. Semiparametric frailty model

On a previous chapter we referred to the semiparametric proportional hazards model. This chapter, discusses the extension of the model, the semiparametric frailty model which constitutes a standard tool to handle multivariate survival data. As mentioned before, the hazard function of a semiparametric frailty model is a product of an unspecified baseline hazard and a parametric function of the covariates that include also the frailty term as a factor. However the direct maximisation of the marginal likelihood (3.2.3) is not possible as in the parametric case, therefore we need to implement a different process in order to get the estimators for ξ, θ and β.

In most of the cases, the Expectation–Maximisation algorithm (EM algorithm) is used as an estimation process in the Frailty models. The EM algorithm approaches the frailty term $u_i$ as missing data (Therneau & Grambsch, 2000) and it involves a repeating combination of two steps, the step of estimation and the step of maximisation. In the beginning, the algorithm obtains the expected values of the frailties, conditional on the observed information and new estimates of the parameters are subsequently obtained, based on the likelihood, given those expected values (Duchateau & Janssen, 2008). Finally the algorithm continues by performing these two steps iteratively. The drawback of this procedure is that the direct estimation of the hazard function is not possible since the produced distribution estimator is discrete (Rondeau, 2003).

## 3.3. Computational methods of estimations

A comprehensive presentation of the EM algorithm, as described by Duchateau and Janssen (2008) is described bellow. The algorithm uses the full data loglikelihood which is given by

$$l_{full}\left(\lambda_0(.),\theta,\beta\right)=\log f\left(z,u\middle|\lambda_0(.),\theta,\beta\right)$$

$$logf\left(z\middle|\lambda_0(.),\beta,u\right)+logf\left(u\middle|\theta\right)=\log_{full},1\left(\lambda_0(.),\beta\right)+l_{full},2\left(\theta\right) \quad (3.4.1)$$

where the conditional on the frailties loglikelihood of z, is $l_{full},1\left(\lambda_0(.),\beta\right)$ given by

$$l_{full},1\left(\lambda_0(.),\beta\right)=\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left[\delta_{ij}\log\left(\lambda_0\left(y_{ij}\right)u_i\exp\left(x_{ij}^t\beta\right)\right)-\Lambda_0\left(y_{ij}\right)u_i\exp\left(x_{ij}^t\beta\right)\right] \quad (3.4.2)$$

One can see that the conditional loglikelihood is only a function of the β and the unspecified baseline hazard function. Finally the second part of the expression (3.4.1) is given by:

$$l_{full},2\left(\theta\right)=\sum_{i=1}^{s}\log f_u\left(u_i\right) \quad (3.4.3)$$

The *maximisation step* uses the first part $l_{full},1\left(\lambda_0(.),\beta\right)$ and the second part of $l_{full},2\left(\theta\right)$ the expression in order to estimate β and θ respectively. Then, the initial form of the expression is profiled to a partial loglikelihood by regarding the frailty term as a fixed offset. So we get:

$$l_{part,1}\left(\beta\right)=\sum_{i=1}^{s}\sum_{j=1}^{n_i}\delta_{ij}\left[\log u_i+x_{ij}^t\beta-\log\left(\sum_{qw\in R\left(y_{ij}\right)}u_q\exp\left(x_{qw}^t\beta\right)\right)\right] \quad (3.4.4)$$

The next step (as every next iteration step k) is for the $u_i$'s and their logarithms to be replaced by the current expected values $E_{(k)}$ ($U_i$) and $E_{(k)}$ ( log $U_i$ ) resulting to the form (3.4.5), which is used so as the new estimates of $\beta^{(\kappa)}$ can be obtained.

$$l_{part,1}\left(\beta\right)=\sum_{i=1}^{s}\sum_{j=1}^{n_i}\delta_{ij}\left[E_{\left(k\right)}\left(\log U_i\right)+x_{ij}^t\beta-\log\left(\sum_{qw\in R\left(y_{ij}\right)}E_{\left(k\right)}U_q\exp\left(x_{qw}^t\beta\right)\right)\right] \tag{3.4.5}$$

Once the estimates of $\beta^{(k)}$ are obtained, estimates for the $\theta^{(\kappa)}$ follow immediately after by implementing the $l_{full},2\left(\theta\right)$ where the $u_i$'s and log $u_i$'s are also replaced once again by the current expected values of iteration step k, as in the expression (3.4.5). The current values of $\beta^{(k)}$ are then used so that the Nelson – Aalen estimator of the cumulative baseline hazard and baseline hazard with the frailties are derived

$$\Lambda_0^{\left(k\right)}\left(t\right)=\sum_{y_{\left(l\right)}\leqslant t}\lambda_{0l}^{\left(k\right)} \tag{3.4.6}$$

and

$$\lambda_{0l}^{\left(k\right)}=\frac{N_{\left(l\right)}}{\sum E_{\left(k\right)}\left(U_q\right)\exp\left(x_{qw}^t\beta^{\left(k\right)}\right)} \tag{3.4.7}$$

So finally we get

$$\Lambda_0^{\left(k\right)}\left(t\right)=\sum_{y_{\left(l\right)}\leqslant t}\frac{N_{\left(l\right)}}{\sum E_{\left(k\right)}\left(U_q\right)\exp\left(x_{qw}^t\beta^{\left(k\right)}\right)} \tag{3.4.8}$$

where $y_{\left(1\right)}<\ldots<y_{\left(r\right)}$ are the times of the events and $N_{(l)}$ is the number of events at time $y_{(l)}$, l=1,...,r.

In addition, R(y$_{(l)}$) is the risk set of subjects at time y$_{(l).}$

The expression (3.4.8) is then used in the expectation step. An assumption is made, that the current estimates at iteration step k are given by $\zeta^{(k)}=\left(\lambda_0^{(k)}(.),\theta^{(k)},\beta^{(k)}\right)$ , which in its turn is implemented so that the expectation $E_{(k+1)}(U_i)=E_{(\zeta)}(U_i|z)$ is obtained. However, the conditional distribution of f$_U$(u$_i$ | **z**) is also needed in this process. Therefore, we take the conditional likelihood for the i-th cluster L$_i$ (ξ,β | u$_i$) (3.a) and the marginal likelihood L$_{\text{marg, I}}$ (ζ) for the i-th cluster (3.2.2) and based on the Bayes theorem, we can write this conditional distribution of the frailties as

$$f_U\left(u_i\middle|z\right)=\frac{L_i\left(\lambda_0(.),\beta\middle|u_i\right)f_U\left(u_i\right)}{L_{marg,i}\left(\lambda_0(.),\theta,\beta\right)}$$

$$\frac{u^{d_i+1/\theta-1}\exp\left(-u_i\left(1/\theta+H_{X_i,c}\left(y_i\right)\right)\right)\left(1/\theta+H_{X_i,c}\left(y_i\right)\right)^{d_i+1/\theta}}{\Gamma\left(d_i+1/\theta\right)} \qquad (3.4.9)$$

Where $H_{\chi_i,c}(y_i)=\sum H_{x_{ij},c}(y_{ij})$ corresponds to a gamma density with parameters (d$_i$ + 1/θ) and $\left(1/\theta+H_{\chi_i,c}(y_i)\right)$ . Hence the expected value $E_{(k+1)}(U_i)$ is given by

$$E_{(k+1)}\left(U_i\right)=\frac{\left(d_i+1/\theta^{(\kappa)}\right)}{1/\theta^{(\kappa)}+H_{X_i,c}^{(k)}\left(y_i\right)} \qquad (3.4.10)$$

Following the same process and since the conditional distribution of log U$_i$ is a loggamma, the expected value of the log U$_i$ is given by

$$E_{(k+1)}\left(\log U_i\right)=\psi\left(d_i+1/\theta^{(\kappa)}\right)-\log\left(1/\theta^{(\kappa)}+H_{X_i,c}^{(k)}\left(y_i\right)\right) \qquad (3.4.11)$$

with ψ denoting the second derivative of the logarithm of the gamma function.

The summary of the E-M algorithm follows

- On the initial step, the algorithm uses a standard Cox regression model to obtain the initial estimates of β and $\Lambda_0$ from the expressions (3.4.5) and (3.4.8) respectively, by setting the frailty term equal to 1 (i.e., θ=0)
- Then by using the current values of $\theta^{(k-1)}$, $\beta^{(k-1)}$ and $\Lambda_0^{(k-1)}$ we obtain the expected values $E_{(k)}(U_i)$ and $E_{(k)}(\log U_i)$
- The next step is to maximize the $l_{part,1}$ (β) (3.4.5) and $l_{full,2}$(θ) (3.4.3) in order to update the current values $\theta^{(k-1)}$, $\beta^{(k-1)}$ and obtain the values $\theta^{(k)}$, $\beta^{(\kappa)}$
- Next, the algorithm iterates between these two previous steps.
- At the final iteration step k, the current values of $\zeta^{(k)}=(\lambda_0^{(k)}(\ .\ ),\ \theta^{(k)},\ \beta^{(k)})$ and $\Lambda_0$ are inserted in the marginal log-likelihood

$$l_{marg}(\zeta)=\sum_{i=1}^{s}\Big[d_i\log\theta-\log\Gamma(1/\theta)+\log\Gamma(1/\theta+d_i)$$
$$-(1/\theta+d_i)\log\Big(1+\theta\sum_{j=1}^{n_i}\Lambda_0(t_{ij})\exp(\beta z_{ij}^{t})\Big)+\sum_{j=1}^{n_i}\delta_{ij}\big(\beta z_{ij}^{t}+\log\lambda_0(t_{ij})\big)\Big]$$

to obtain its value at iteration level k.

Finally, the algorithm reaches convergence when the absolute difference between $l_{marg}\left(\lambda_0^{(k-1)}(.),\theta^{(k-1)},\beta^{(k-1)}\right)$ and $l_{marg}\left(\lambda_0^{(k)}(.),\theta^{(k)},\beta^{(k)}\right)$ is smaller than a preset value ε. Consequently the observed information matrix, I, for $\hat{\theta}$ and $\hat{\beta}$ is obtained by calculation based on the observable log-likelihood using the joint distribution of ( $Y_{ij}$ , $I_{ij}$ ) (Klein, 1992).

The drawback of this procedure is that the direct estimation of the hazard function is not possible since the produced distribution estimator is discrete (Rondeau, 2003). An alternative approach to the EM algorithm is the Penalised Partial Likelihood Estimation (PPL). In addition, PPL converges normally faster than EM while producing both point and variance estimates for the parameters of interest (Hanagal, 2011). However, Duchateau &

Janssen (2008) showed that, in the context of the semiparametric gamma frailty model, EM algorithm yields the same results as PPL.

## 3.4. Multiple events per subject

Multiple event-time data are gaining more and more attention in the survival analysis during the recent years. Multiple events can be classified according to whether they are different types of events or recurrences of the same types of events. According to this classification we refer to the multiple events as competing risks or recurrent events respectively. We will focus on the latter case where subjects present the event of interest more than one time while being under observation.

The main interest in this case is whether there is a correlation of the recurrent events within a subject. There are two main approaches to modeling this correlation that have gathered a lot of attention and both of the approaches comprise extensions of the proportional hazards models.

The first approach is the Variance-corrected models which do not include the dependence of event time, but instead they adjust for the additional correlation by implementing the covariance matrix of the estimators. The term V*ariance-corrected models* emanates from the fact that the standard variance estimate is replaced by another corrected estimate that accounts for possible correlations.

The Variance-corrected models are extensions of the ordinary Cox model, where the estimate of variance for the vector of coefficients treats every observation independent. However, in multiple event-time data this assumption is not valid since subjects may contribute more than one observation. Therefore, we need a corrected variance estimator that treat observations in their clustered form. Lin and Wei (1989) proposed a robust correction for the naive estimator of the ordinary cox model

$$I^{-1} = -\partial^2 \log L\left(\beta\right) / \partial\beta\partial\beta'$$

The proposed robust or sandwich estimator is given by

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t < T \leqslant t + \Delta t)$$

where U is a n x p matrix of efficient score residuals and D is the n x p vector of leverage residuals that are taken from the differences of the estimated β if each observation i is taken out of the data set. When observations are clustered in m independent clusters ($G_1$, $G_2$, … , $G_m$) and not independent as the above formula assumes, then the robust covariance matrix is given by

$$V = I^{-1} G' G I^{-1}$$

where G is a m x p matrix of the cluster efficient score residuals.

The second approach is the shared frailty models where this correlation is modeled by a random effect that is assumed to be independent conditional on the per-subject coefficient. We will extend to the shared frailty models in the chapter that follows.

Another classification, which may overlap on some level with the previous classification and is commonly used for the multiple events per-subject data, is according to whether events have a natural order of occurring or not. If events have a natural order then we refer to them as Ordered failure events and if not, Unordered failure events.

In most of the cases, unordered failure events include different events by the same subject which are called competing risks and subjects can be under risk for multiple events simultaneously. Same type of events can be treated as unordered failure events, such as in the case of paired data where two organs of the same subject, like eyes, can be assigned to two different arms of treatment for the same risk. However, this is not common since we are restricting the events to have no order, even though in reality there is a natural order.

Ordered events are most commonly used for the same type of events. The same approaches that are used for the correlation of the events which were mentioned before can also be used for the assessment of these events. Different timescales can be used as well as counting process formation to notate this natural order.

### 3.4.1. Shared frailty models for recurrent events

As noted above, in the context of recurrent events, the term 'shared' refers to a single individual and denotes the fact that is shared over time, meaning that there is only one individual sharing each value of the random effect (Hougaard, 2000). Therefore what used to be group variation, now is variation between individuals described by the random variable (frailty term) and the variation that is described by the hazard function is now the within-subject variation.

In addition to the assumption of all times being independent given the values of the frailties, it is also assumed that the individuals are independent. It should also be clear that whenever there is a single type of event, as it is in our case, then the event times are ordered with $0 < T_1 < T_2 < \dots$, meaning that there is no possibility that two succeeding events within a subject coincide, hence there are no ties. In that sense, we need to implement a more complex structure when describing recurrent event data, in order to keep track of the event order within every subject.

## 3.5. Time-scales for recurrent events

In chapter (2.3) we mentioned the different time-scales that can be used in order to represent the data. Likewise, recurrent event data can be represented in different ways based on a different time-scale each time. Kelly and Lim (2000) used three different timescales in the five predominant Cox-based models for recurrent events (Andersen and

Gill (AG), Prentice Williams and Peterson gap time (PWP-GT) and elapsed time (PWP-CP, CP denotes the counting process form of the data), Lee, Wei and Amato (LWA), and Wei, Lin and Weissfeld (WLW)). Their application included the gap, the calendar and the total-time timescales. A visual representation of the timescales that were used can be seen in figure 3.1. As far as the total-time timescale is concerned, one can see that it has limited appeal since subjects would be at risk for all of their events at the initiation of monitoring even though some events need previous events in order to take place. Therefore, we won't refer to it for the remainder of this thesis.

Kelly and Lim (2000) argued that adopting a different timescale determines the nature of the model that is used. In the Gap timescale setting, the model is conditional since the individual under observation is at risk for the k$^{th}$ event if only the (k-1)$^{th}$ has already occurred, while in the total time-scale the model is marginal since being at risk for the k$^{th}$ event does not depend on the previous events. They showed that these models, which depend on robust variance estimate to adjust for misclassifications, do not account adequately for the potential correlation that occurs in situations where a subject has multiple events, the so called within-subject correlation. In order to face this problem, they proposed the use of the frailty models, where this correlation is modelled by using a random effect.

The timescale that is most commonly used in the context of frailty models is the gap timescale where the subject starts again at time 0 after an event and the time at risk for a subject corresponds to the time it takes for the next event to occur. In other words, the clock restarts after every event. Under the total time, the time at risk is measured for every event from time 0 (entry to the study) till the occurrence of the event regardless if other events have occurred meanwhile.

Let us now clarify how we define the different time-scales by introducing the concept of risk intervals. A risk interval is the period that an individual is at risk of having the event of interest. Every risk interval is interrupted either by an event or by a period that the subject was not under observation. For every interval there is a starting and an end point. The starting point can be the entry to the study, the moment after an event or censoring has occurred and the end point is the time of event occurrence or the time of censoring.

Duchateau et al. (2003) extend the previous work by implementing frailty models to account for the within-subject correlation in a study of recurrent asthma attacks by also taking into account the time that an asthma attack lasts. Taking into account how much an asthma attack (or generally the event of interest) lasts introduce an accessory problem to the analysis since a way must be found to address the issue of what should be done with the time that the subject was not at risk either due to the duration of the event or because of censoring.

Adopting one of these timescales determines the hazard function differently. So, let's assume that there are N subjects in total (i=1,2,...,N) and that for every subject i there are $r_i$ risk intervals. This time the risk intervals differ from the definition of Kelly and Lim since the time that an event or censoring last is taken into account. Consequently, for every $r_i$ risk interval there is a $t_{i\,j1}$ starting point of the j-th risk period and $t_{i\,j2}$ end of the j-th period. In the Calendar time the risk interval for a specific event is the time from the end of the previous event to the start of the next event.

This way, under the Calendar timescale model the complete information that we have for the i-th subject can be summarised by the $r_i$ triplets

$$\left(\left(t_{i11}, t_{i12}, \delta_{i1}\right), \ldots, \left(t_{ir_i1}, t_{ir_i2}, \delta_{ir_i}\right)\right) \tag{3.3.1}$$

so the hazard function is given by

$$\lambda_i(t) = \begin{cases} \lambda_0(t)\, U_i \exp(\beta z_i)\, for\, t_{ij1} \leq t \leq t_{ij2},\, j=1,\ldots,r_i \\ 0\, otherwise \end{cases} \quad (3.3.2)$$

Then the corresponding likelihood is given by

$$L(\beta) = \prod_{i=1}^{N} \prod_{j=1}^{r_i} \lambda_i\left(t_{ij2}\right)^{\delta_{ij}} \exp - \Lambda_i\left(t_{ij1}, t_{ij2}\right) \quad (3.3.3)$$

with the cumulative hazard given by

$$\Lambda_i\left(t_{ij1}, t_{ij2}\right) = \int_{t_{ij1}}^{t_{ij2}} \lambda_i(t)\, dt \quad (3.4.3)$$

Under the gap timescale the information that we have can be summarised in the triplets

$$\left(\left(t_{i12} - t_{i11}, \delta_{i1}\right), \ldots, \left(t_{ir_i2} - t_{ir_i1}, \delta_{ir_i}\right)\right) \quad (3.3.5)$$

so the hazard function becomes

$$\lambda_i(t) = \begin{cases} \lambda_0\left(t - t_{ij1}\right) U_i \exp\left(-\beta z_i\right) for\, t_{ij1} \leq t \leq t_{ij2},\, j=1,\ldots,r_i \\ 0\, otherwise \end{cases} \quad (3.3.6)$$

The likelihood function is the same as in the calendar timescale setting but now the interpretation of the hazard function λ$_i$(.) is different as in the gap timescale the ordering of events is ignored. This is succeeded by resetting the start of risk period j at time 0 while using the gap timescale, whereas in the calendar timescale the start of risk period j corresponds to the actual time since the subject entered the study. However, in both cases, the length of each risk period remains the same.

These two timescales can be adopted in the case of the semiparametric frailty model where we get two different expressions for the likelihood function in respect to the risk sets that are used. The partial likelihood function under the Calendar timescale can be written as:

$$L(\beta)=\prod_{i=1}^{N}\prod_{j=0}^{r_i}\left\{\frac{U_i\exp(\beta z_i)}{\sum_{k=1}^{N} Y_k(t_{ij2})U_k\exp(\beta z_k)}\right\}^{\delta_{ij}} \tag{3.3.7}$$

where $Y_k\,(t_{i\,j2})$ is an indicator for the k-th subject being (or not) at risk, given by

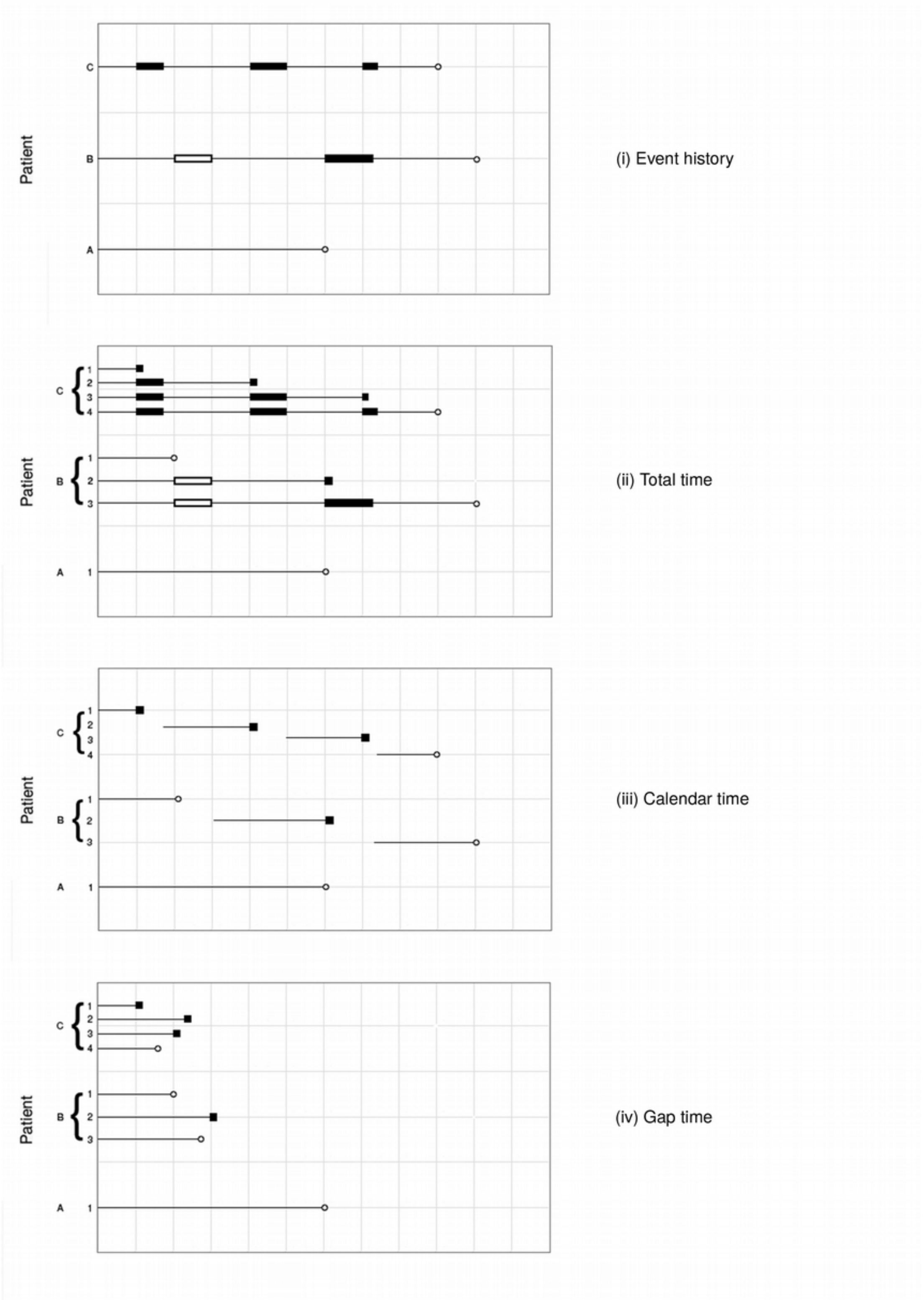
$$Y_k(t_{ij2})=\begin{cases}1\ \mathit{if}\ k-\mathit{th\ subject\ is\ at\ risk\ at\ time}\ t_{i\,j2}\\ 0\ \mathit{otherwise}\end{cases} \tag{3.3.8}$$

Following the same notation, the partial likelihood function under the Gap timescale is given by

$$L(\beta)=\prod_{i=1}^{N}\prod_{j=0}^{r_i}\left\{\frac{U_i\exp(\beta z_i)}{\sum_{k=1}^{N}\sum_{l=0}^{r_k} Y_{kl}(t_{ij2})\,U_k\exp(\beta z_k)}\right\}^{\delta_{ij}} \tag{3.3.9}$$

with the indicator for the risk being $Y_{k\,l}\,(t_{i\,j2})$

$$Y_{kl}(t_{ij2})=\begin{cases}1\ \mathit{if}\ t_{kl2}-t_{kl1}\geq t_{ij2}-t_{ij1}\\ 0\ \mathit{otherwise}\end{cases} \tag{3.3.10}$$



(i) Event history

(ii) Total time

(iii) Calendar time

(iv) Gap time

**Figure 3.1:** *Illustration of (i) Event history along with (ii) total time (iii) Calendar time and (iv) Gap time of 3 patients with recurrent events (* ■ *denotes events and* O *censoring)*

Something that comes into question though, is the reasoning to use a different timescale in the first place. Each time we adopt a different timescale, we get an answer to a different research question and every question brings out a different aspect of the data. In the Calendar timescale model our main concern is the evolution of how often the event of interest occurs since the entry to the study while in the Gap timescale model what seems to be more interesting is the effect that an event has on the rate of the occurrence of a subsequent event.

In addition, we can often evaluate the need for using different timescales prior to the analysis. If we have a basic idea about the evolution of the event of interest and how susceptible its rate is to change over time, we can decide if we need to implement more timescales. That is, if we do not expect a radical change in the rate of the recurrent event then we can simply use the gap timescale. In reverse, if we do expect a radical change then it is better to fit more models (Duchateau, 2003).

## 3.6. Variance-corrected Models for recurrent events

The concept of an event occurring to a subject more than once during the observational period is an extension of the single event model. In this chapter we are going to present in more details the AG, PWP-CP and PWP-GT models that we mentioned in a previous chapter, and how different timescales can be used. In the variance-corrected model approach, the correlation of event times is not included in the model as it does in frailty models; instead the covariance matrix of the estimators is adjusted to account for the additional dependence.

The simplest modeling approach based on the Variance corrected models is the AG model, proposed by Andersen & Gill (1982). In this model, all subjects are at risk for each event at all times. This implies that all subjects share a common baseline rate function.

Having a common baseline function for all events can be succeeded by using counting process in the definition of the risk intervals.

The PWP-CP and PWP-GT models are also called conditional models. The models are conditional in the sense that a subject cannot be at risk for a second event until the first event has occurred and so on. Contrary to the AG model, conditional models stratify data according to each event, this way, they allow baseline hazards to vary with each event. However, the PWP-CP model is essentially an AG model with event-specific baseline hazards by stratification. The PWP-GT model is a little different as it uses the gap timescale instead of counting process.

Even though two different timescales can be used, let us clarify that, both of these “different” models focus on the survival time between two gaps. The difference of the models comes up in the definition of the risk sets. In the gap-timescale, the clock resets at time 0 after each event as we have mentioned before, while in the calendar time, uses the actual times from study entry. The hazard function under the PWP-GT model for the i$^{th}$ subject and k$^{th}$ event, under the assumption that all the covariates at the start of the study are fixed, is given by:

$$\lambda^{(k)}\left(t; Z_i^{(k)}\right) = \lambda_0^{(k)}(t)\exp\left(Z_i^{(k)}\beta\right)$$

The corresponding hazard function under the PWP-CP is given by:

$$\lambda^{(k)}\left(t; Z_i^{(k)}\right) = \lambda_0^{(k)}\left(t_k - t_{k-1}\right)\exp\left(Z_i^{(k)}\beta\right)$$

We should also make clear that the conditional model with the gap timescale takes into account the natural order of the recurrent event contrary to the simple frailty model where the order is ignored while using the gap timescale. In order to succeed that, a variable with the ascending number of the event is included to the dataset and is used to stratify the data, as we mentioned before.

Suppose now that we have a study with n patients that lasted 20 months that focuses on multiple subsequent infections. Each patient after having experienced the event of interest (e.g. infection), is treated for it and returns to the study once being healed. While being treated, subjects are assumed not to be at risk, therefore after each event there is a risk-free period for every subject. Out of these n patients we will focus on three independent patients for the illustration of the dataset for the variance corrected models. The first patient experienced the event of interest twice, one at month 5 and the other at month 9 and after that he was followed for 7 additional months till the study ended and he was censored. After his first event, he was treated for 2 months before he returned to being at risk while after his second event the healing period, hence the risk-free period, lasted 4 months. The second patient experienced the event three times, at 8, 16 and 19 month respectively and never returned for follow-up observation after the last infection. His healing process lasted 3 months after event one and 1 month after the second event. Finally, the third patient experienced the event once at month 4 and returned for follow up at month 9 until he was censored.

In table 3.1 we see the data layout of the three hypothetical subjects for the three variance-corrected models. One can see that there is no difference in the risk periods for each of the three hypothetical patients since each patient is at risk for the same time, however there are differences in the risk sets. Let us take now the risk set for the second event for instance. We notice that under the calendar timescale the first patient enters the risk set at month 7 and exits at month 9, the second patient enters at 11 and exits at 16, while the third patient enters at 9 and exits at 20 when the study ends and contributes to the partial likelihood a censored survival time. Under the gap timescale all three patients enter at time 0 and they are all present until 4 when patient 3 exits. Additionally, the duration of healing which takes place immediately after each event, is incorporated since each subject is not at risk for the next event immediately after the preceding event but it becomes at risk at the end of the healing process.

Table 3.1: Variance-corrected models for recurrent events, data layout for three hypothetical patients

| | Patient 1 | | | Patient 2 | | | Patient 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | time interval | Event | Stratum | time interval | Event | Stratum | time interval | Event | Stratum |
| AG | (0,5] | 1 | 1 | (2,8] | 1 | 1 | (0,4] | 1 | 1 |
| | (7,9] | 1 | 1 | (11,16] | 1 | 1 | (9,20] | 0 | 1 |
| | (13,20] | 0 | 1 | (17,19] | 1 | 1 | | | |
| PWP-CP | (0,5] | 1 | 1 | (2,8] | 1 | 1 | (0,4] | 1 | 1 |
| | (7,9] | 1 | 2 | (11,16] | 1 | 2 | (9,20] | 0 | 2 |
| | (13,20] | 0 | 3 | (17,19] | 1 | 3 | | | |
| PWP-GT | (0,5] | 1 | 1 | (0,6] | 1 | 1 | (0,4] | 1 | 1 |
| | (0,2] | 1 | 2 | (0,5] | 1 | 2 | (0,11] | 0 | 2 |
| | (0,7] | 0 | 3 | (0,2] | 1 | 3 | | | |

## 3.7. The conditional shared frailty model

Till now we have mentioned the simple frailty model which incorporates the heterogeneity of the subjects in the frailty term. This model, like the AG model, does not allow the baseline hazard rate to vary by the event number. On the other hand, the variance corrected models that allow for different baseline hazards, do not incorporate the heterogeneity into the estimates themselves, instead they rely on the robust standard errors to assess the heterogeneity and this is why they remain biased (Box-Steffensmeier & De Boef, 2006).

Even though shared frailty models perform well in assessing the heterogeneity of the subjects by the use of the frailty term, they lack in reducing the possible biases due to the dependency of recurrent events. An attempt to incorporate both heterogeneity of the subjects and dependency of the events was made by Cook and Lawless (2007). Their attempt involved a mixture of semiparametric shared gamma-frailty model with Poisson counting process where the subject specific function was given by:

$$\lambda_i\left(t|H_i\left(t\right)\right)=\lim_{\Delta t\to 0}\frac{P\left\{\Delta N_i\left(t\right)=1\,|\,H_i\left(t\right)\right\}}{\Delta t}=u_i\rho_i\left(t\right)$$

where u$_i$ are the random effects and given those effects and covariates, { N$_i$ (t), 0 ≤ t } is a Poisson process with rate u$_i$ ρ$_i$ (t).

The mixture of the semiparametric shared frailty model with the counting process expressed by the stratification by the event is what we are going to simply call conditional frailty model. It combines the incorporation of unobserved heterogeneity by the frailty term and the incorporation of dependency due to recurrent events by stratification. The hazard function for the j$^{th}$ subject and the k$^{th}$ event is given by:

$$\lambda_j^{(k)}\left(t\right)=\lambda_0^{(k)}\left(t_k-t_{k-1}\right)u_j\exp Z_j^{(k)}\beta$$

The corresponding partial likelihood for this model, given the frailties, is given by:

$$L\left(\beta\right)=\prod_{j=1}^{n}\prod_{k=1}^{K}\left\{\frac{U_j\exp\left(\beta Z_i^{(K)}\right)}{\sum_{j=1}^{n}\sum_{k=1}^{K}Y_{jk}u_j\exp\left(\beta Z_j^{(k)}\right)}\right\}^{\delta_j^{(k)}}$$

where δ is the censoring indicator which is equal to 1 if the k$^{th}$ event is observed for the j$^{th}$ subject and 0 otherwise, while Y is the risk indicator, which takes value 1 if the subject is at risk for the k$^{th}$ event and 0 if not.

## 3.8. Martingale Residuals

After fitting the model, its adequacy needs to be assessed and that is crucial in the modeling process. Using the martingale residuals is a well established method for checking the adequacy of the fitted model as it gives us the ability to decide whether the model's prediction of the number of observed events is right (Mazroui, 2016). However, in the presence of censored survival times, visual inspection of the model fitting becomes somehow more complicated than the methods used in the linear regression modeling (Collett, 2003) where errors are assumed to be normally distributed.

Martingale residuals are based on methods known as martingale methods, hence their name. These residuals comprise a refinement of the modified Cox-Snell residuals and can be calculated by using the counting process formulation. Therefore, the basis for the Martingale residuals are the difference between the counting process and the integrated intensity function (Andersen et al, 2012)

$$M_i(t) = N_i(t) - \int_o^t Y_i(s) \exp\left(\beta Z_i(s)\right) d\Lambda_0(s) \tag{3.5.1}$$

If we denote $\hat{\beta}$ the estimator of β by the maximum partial likelihood then the Breslow (1974) estimate of the cumulative hazard is given by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum dN_i(s)}{\sum Y_j(s) \exp\left(\hat{\beta} Z_j(s)\right)} \tag{3.5.2}$$

By implementing this expression in (3.5.1), we define the martingale residual as (Therneau et al. 1990)

$$\widehat{M}_i(t)=N_i(t)-\int_o^t Y_i(s)\exp\left(\hat{\beta}Z_i(s)\right)d\widehat{\Lambda}_0(s) \tag{3.5.3}$$

A simplified form of expression 3.5.3 for the simple semiparametric proportional hazards model is

$$\widehat{M}_i=\delta_i-\widehat{\Lambda}_0(x_i)\exp\left(\hat{\beta}Z_i\right) \tag{3.5.4}$$

where $\delta_i$ is the failure indicator and $x_i$ is the observational time. The residual Mi can be interpreted as the difference between the observed number of events (0 or 1) for subject i between time 0 and $T_i$ , and the expected numbers based on the fitted model. Barlow & Prentice (1988) suggested a graphical use of these residuals in order to assess the goodness-of-fit. It was also suggested that it is often sufficient that residuals are centered about zero with a known scaling.

## 3.9. Akaike information criterion

Akaike information criterion is one of the most commonly used measure of the relative quality when it comes to survival models and in that sense it provides a means for model selection. It was first introduced by Hirotugu Akaike in 1973. Its general form is given by:

$$AIC=-2\log\left(maximized\ likelihood\right)+2p$$

where p is the number of parameters of the model.

The criterion combines the goodness of fit of the model via the computation of its likelihood with the trade-off of the complexity of the model which is penalized by double the number of parameters of the model. It is called criterion and not a test since it does not provide a testing of some null hypothesis but rather an information criterion in the form of a relative estimate of the information lost.

Using the AIC as a process of model selection is a common practice in many occasions. In the framework of recurrent event analysis it has also been used extensively. More specifically, Duchateau et al. (2003) used AIC to select the most appropriate model between fits of different timescales. Additionally, the same process was later used by Ullah et al. (2012) while analyzing recurrent sport injuries in order to evaluate the goodness-of-fit of their fitted models.

# Chapter 4.

# Setting

In this chapter we set the foundation for our analysis, present some of the terms that we are going to use later in the actual analysis and finally declare and provide justification for the assumptions undertaken later on.

## 4.1. Adherence to the treatment and gap definition

In this thesis, we are trying to estimate the gaps in the treatment, and deductively the response to the treatment, among HIV-positive patients in Greece. For this reason, we are trying to investigate if there is any correlation between the occurrence of consecutive gaps in the clinical visits of these patients.

The individuals that are included in the AMACS, along with being monitored, are also undergoing a variety of medical and biochemical examinations as well as receiving treatment for HIV and other possible coinfections. The attendance on those visits is crucial, since the clinical state of a patient can be reevaluated with each appointment he attends and a better course of action concerning their health improvement can be adapted. Therefore, patient attrition pose a great threat to the success of antiretroviral therapy programs and the fight against the epidemic.

With that been said, we need to define what constitutes a gap in attendance to monitoring and treatment appointments, or simply gap in care/treatment, and therefore what is considered event or failure to our analysis.

A patient is categorized as “active” if he returns to a clinic for monitoring within a year since his last clinic encounter; in reverse, a patient is categorized as “non-active” if a year has passed since his last clinic encounter. However being categorized as non-active means that the only information we have is that the gap in care/treatment occurred somewhere within this one year interval. Hence this definition implies that our data are interval censored which makes the construction of the likelihood, where we base our inferences, more complex. A way to overcome this drawback is by assuming a specific moment that each patient “decided” not to attend his next appointment for monitoring or care, within the year of his absence.

A problem arises while trying to define the moment of incidence since there is no consensus in the definition of gap in care. The work of Chi et. al. (2011) is based on the need of setting a standard for studies like ours. Working with a substantially large study population of HIV positive people from more than 100 health facilities, they recommend a threshold of 180 days since the last clinic encounter as a universal definition for gap in care/treatment. Taking in to account this recommendation we define the median (182.625 days) of the one year interval as the moment when the event occurred.

After having defined what constitutes event of interest for our study, the next step is to clarify the at-risk and risk-free periods for the subjects according to what was mentioned in the previous chapter. Under the calendar timescale, a subject’s at-risk period for the first event begins at the initiation of monitoring. Then we run through all registered visits of each patient until a year long gap (or longer) is found. After having identified the date of the last visit before the gap, we add 182.625 days to this date to signify the end of the at-risk period. If a patient has a subsequent visit, then his next at-risk period spans from this visit until he is either censored or committing a second gap. Subsequently, the risk-free period is signified by the in-between time. This process continues till there are no more visits to check.

## 4.2. Definition of study endpoint / censoring

As mentioned before, one of the most crucial concepts in Survival analysis is censoring. Therefore we need to define the study endpoint which is the moment that patients who are still monitored, will be right censored.

Since AMACS is an ongoing study, data on HIV-positive patients are still collected as the moment we speak, therefore there is no actual endpoint of data collection. Additionally, AMACS collaborates with 13 clinics and hospitals scattered in Greece, making the universal definition of study endpoint ever more complex.

Taking these facts into account, we defined thirteen distinct endpoints instead of a universal definition of our study endpoint. Each clinic or hospital that collaborates with AMACS has a different endpoint that was defined by the latest date that they provided data of their patients; thereby, each patient, depending on the clinic or hospital that they attend, got his censoring time due to administrative censoring.

However, following this process, led to another problem since many of the patients have visited more than one clinic or hospital within their observational time interval. In order to deal with this problem, we used as the reference clinic or hospital the one that the subject has visited most times.

In the second chapter of the thesis we referred to the main possible sources of right censoring, which are withdrawal from the study, loss to follow up and competing risks. However, since the event of interest in our analysis is the gap in the attendance and not death, we have to redefine the censoring reasons. Consequently, in our data, there are two reasons that a subject is censored. The first is when the database closes (administrative censoring) and the second is if a patient dies, indifferent to the cause of death.

## 4.3. Longitudinal measurements

In studies like ours, one is interested in estimating the evolution of an outcome variable adjusted for a set of explanatory variables. These variables can be either qualitative like arm of treatment, sex and origin or quantitative like age, virus load or CD4 count. Many models proposed in literature include these variables in the univariate form, where the values of the measurements are determined at the moment that a subject enters the study. However, in a longitudinal study like ours, it is irrational to assume that variables related to disease progression (like CD4 counts and viral load) stay constant throughout the whole period of observation, especially since there are subjects that are monitored for almost two decades. Thereby, the hazard function is more depended on the updated values of those time-depended covariates rather than on their value at the initiation of observation.

While it is rational to include covariates in the model that are time depended, we need to take into account the nature of these covariates before doing so, since the model may become more complicated. In addition, model over-fitting is another problem that arises when time-depended covariates are included in a model. On the other hand, some variables, such as the subject’s age, that conceptually makes sense to treat them as time-depended covariates, including them as such would not affect their estimated coefficients since the evolution of the subject's age is absorbed into the baseline hazard function (Hosmer & Lemeshow, 2008). As long as age is included into the model in linear form and not in a more complex way such as quadratic form or spline, the effect remains the same (Therneau & Grambsch, 2000) irrespectively if it is treated as time-constant or time-dependent. The reason the estimated coefficients are the same in both cases is that the internal computations which are used for the models depend only on hazard ratios. Let us take for instance the age at enrollment and current age for the illustration of the previous sentence. Additionally, let us examine the hazard ratio at time t for two subjects, subject “a” and subject “b”. Their age at enrollment would be age$_a$ and age$_b$ while their current age at time t would be age$_a$ + t and age$_b$ + t respectively. Their hazard ratio is given by:

$$HR=\frac{\lambda_0(t)\exp\left(\beta\left(age_a+t\right)\right)}{\lambda_0(t)\exp\left(\beta\left(age_b+t\right)\right)}=\frac{\exp\left(\beta\, age_a\right)}{\exp\left(\beta\, age_b\right)}$$

To generalize the simple Cox model hazard function (2.6.2) by including time-depended covariates is not difficult and is given by

$$\lambda\left(t,z(t),\beta\right)=\lambda_0(t)\exp\left(z_1^T(t)\beta_1+z_2^T\beta_2\right) \tag{4.1}$$

It is very important though, to realize that this is no longer a proportional hazards model as it would be if we only included constant covariates. However the proportionality could be implied. Supposing that we had a model including the measurement of the CD4 cell count at time t (which is time-depended covariate) and sex (which is constant covariate). If we would denote z$_1$(t) and z$_2$, cd4 count and sex respectively the hazard function of the model would be given by

$$\lambda\left(t,z(t),\beta\right)=\lambda_0(t)\exp\left(z_1(t)\beta_1+z_2\beta_2\right)$$

Consequently, the hazard ratio for sex would be given by

$$HR\left(t,z_2=1,z_1=0\middle|z_1(t)\right)=\frac{\lambda_0(t)\exp\left(z_1(t)\beta_1+\beta_2\right)}{\lambda_0(t)\exp\left(z_1(t)\beta_1\right)}=\exp\left(\beta_2\right)$$

One can see that this hazard ratio does not depend on time. In that sense we can still call it proportional to a function of time (Hosmer & Lemeshow, 2008).

Likewise with (4.1), the generalized form of the frailty model hazard function (3.1) with time-depended covariates (Huang, 2010), is given by

$$\lambda_i(t)=\lambda_0(t)u_i\exp\left(z_{1i}(t)^T\beta_1+z_{2i}^T\beta_2\right) \tag{4.2}$$

## 4.3.1. Martingale Residuals per subject

In the previous section we referred to the multiple observations for a subject due to time-depended covariates. In the framework of shared frailty models, where we also get multiple observations per subject since every patient comprises a “group”, the subject of how to calculate martingale residuals arises. It seems rational that martingale residuals should also be grouped on the level of every patient. This grouping, as Commenges (2000) explains, implies a way of smoothing the graph of residuals by summing them up for every group.

Suppose that we have n subjects in our dataset. If we now denote m the count of observations, we have m > n observations. Most of the statistical programs will return m residuals, one per observation. According with what was said in the previous paragraph, the per-subject residual for a subject i, would be the sum of the residuals for his multiple observations.

Once the summed martingale residuals are calculated, one per subject and totally n, we can plot them versus the linear predictor or any of the individual covariates. Plotting these residuals however is complicated by the fact that we have m observations and n residuals. To resolve this problem, one can use a constant covariate that is measured at the enrollment of the subject to plot the residuals against it.

# Chapter 5.

# Application in simulated data

In this chapter we apply the frailty models based on the methods we mentioned in the previous chapters in simulated data and present the results.

## 5.1. Simulated data

We use simulated data so that we can compare the conditional frailty model with the other Variance-corrected models as well with the simple semiparametric Cox model and its extension the semiparametric gamma frailty model. We apply the fits in both gap and calendar timescales.

For that reason, we generated in R, two pairs of 1000 data sets for recurrent events for two corresponding scenarios as seen below:

- Scenario A: Heterogeneity among the subjects

- Scenario B: No heterogeneity

The first scenario includes observations of subjects that are heterogeneous in the sense of some subjects being more prone than others to the (re)occurrence of the event of interest for some unknown or unmeasured reason. In the second scenario the groups of observations by each subject is assumed to be homogeneous, meaning that there is no heterogeneity between subjects. In all cases the size of the simulated sample is N=1000.

The data for each subject j are simulated by following the multiplicative intensity function described by Andersen & Gill (1982). More specifically, the intensity process that is used for the generation of data is given by

$$Y_j(t)\lambda_0(t)u_j\exp(\beta X_j)$$

where $Y_j(t)$ is a predictable process which is equal to 1 when subject j is under risk. Within-subject correlation (or between subject heterogeneity) is induced to the data by the frailty term u which follows a gamma distribution with E(u)=1 and V(u) = θ. Censoring times follow a Uniform distribution in the interval (0,10) (time in years).

Disjointed risk intervals are included in the simulated data, since a subject can be at no risk after the (re)occurrence of an event. The risk free period of a subject follow a uniform distribution in the interval of (0.5 , 0.8). The specific values were chosen in order to resemble our real life data that we are going to analyze in the next chapter.

We assume that the heterogeneity between subjects are quite large and for that reason we picked θ = 0.7, as the value for the variance of the frailty term. In addition, we included an explanatory covariate that follows a Binomial distribution, with the probability p=0.4 for the covariate x being equal to one. We assume that the effect of the covariate is also quite large so we pick β = 0.8 for the true covariate effect that we are going to compare to the results of the models. Since x is a boolean variable with values 0 and 1, we assume that it represents smoking status of the subject, with 0 corresponding to non-smoker.

## 5.2. Results of simulated data

We estimated all models in R, by the use of the COXPH procedure in the package 'survival'. In the cases when a model did not converge because of the imbalance in the groups, the loop algorithm repeated the specific step until convergence was succeeded.

Table 5.1 shows the distribution of the estimated variance θ for the 1000 simulated data under both timescales, both scenarios and finally both models (simple frailty and conditional frailty) as well. In all cases, it seems that the simple frailty model performs more adequately in assessing the frailty term. In scenario A where the variance of the frailty term is set to 0.7, we observe that under both timescales the frailty fit never estimated a no-frailty model whereas the conditional frailty fit estimated 40.6% and 80.2% of the time no-frailty model under the gap and calendar timescale respectively. Additionally, the estimated frailty variance by the simple frailty fit is far more closely clustered around the true value of θ=0.7

- Scenario A (Heterogeneity)

As been said before, heterogeneity among subjects means that some subjects are more frail to present the event of interest compared to the rest. That will result the risk sets for the larger number of events to be dominated by more frail subjects than the risk sets of the earlier events while using the models that are stratified by the natural order of events. The simple Cox models performed quite fairly in both timescales as evaluated by the coverage rate of the estimated β, which was 88% (table 5.2). However coverage improved in the marginal models since the confidence intervals are computed using the Robust estimate of the standard error, and thus were wider. This comes natural since variance corrected models, correct for the unobserved heterogeneity through those robust estimates and not during the estimation of the coefficients as frailty and conditional frailty models do.

However, the good performance of the AG model should be mostly credited to the limitation of the simulation method as the simulated datasets were generated using the Andersen & Gill's multiplicative intensity function.

As for the conditional models, both PWP-GT and PWP-CP seem to underestimate the effect of the covariate in the (re)occurrence of the event. In both cases, the mean estimated effect of the covariate was approximately 0.67 while the true effect was set equal to 0.8. In addition, in both timescales, these models had the lowest coverage rate of the true coefficient by the estimated 95% CI's along with the conditional frailty model in the calendar timescale.

The simple frailty models seem to perform better in both timescales while they produce more accurate estimates as well. The coverage rate of true β by the estimated CI's, is the largest compared to the other models since in both cases approaches the 95% of the cases. Additionally, in both timescales, the rejection rate of the estimated θ (the cases where the model results in non statistically significant estimate for the frailty term) is less than 4% and specifically under the calendar timescale where only 1.3% of the cases θ is rejected as seen in table 5.2.

As far as the conditional frailty model is concerned, one can see that the results between the two timescales vary a lot in respect of the mean estimated θ, coverage rate of true β and rejection rate of θ as well. However these differences between the two fits do not exactly mean that one performs better than the other, since both perform poor in all categories mentioned before. Indicative of the poor accuracy for these models is the fact that under both timescales the rejection rate of the variance θ is high. More specifically, in only 20% of the cases under the calendar timescale, θ seems to be playing a significant role in assessing the unobserved heterogeneity. The difference between the rates of the models could be that the conditional frailty model under the calendar timescale is over fitted since both timescale and stratification account for the natural order of the events.

- Scenario B (No heterogeneity)

In scenario B, where no heterogeneity among subjects is incorporated in the simulation process, we observe that all models under both timescales are performing very well with respect to the estimation of the true covariate effect. In all cases the coverage rate is approximating 90%, while the simple frailty models had the highest coverage rate with values close to 95%. Only the two conditional models seem to overestimate the effect of the covariate, however the extent of the overestimation remains low and should probably be credited to the cases where θ is not rejected by the analysis.

Simple shared frailty models seem to be rejecting correctly the variance of the frailty term in all simulations giving a very satisfying rate equal to one. On the other hand conditional frailty models falsely do not reject the variance of the frailty term in 6% and 8% of the cases, under the gap and calendar time respectively.

Figure 5.1 displays the densities of the estimated coefficients by the models. The curves that correspond to the PWP-GT, PWP-CP and both conditional frailty models for the two timescales are shifted to the left indicative of their poor performance in estimating the true β which is also reflected in table 5.2. Once again both simple frailty models seem to be practically unbiased.

Based in all previous comments, the simple frailty model is recommended in both cases where heterogeneity is present which is what is of more interest. In addition, the frailty model under the calendar timescale seems to perform better with regard to rejection rate since it rejects falsely θ in less cases compared to the same model under the gap timescale. Therefore, if the researcher has no particular interest in studying how the hazard rate of the recurrent event evolves after an event has taken place then the analysis can be solely based on the calendar timescale.

As mentioned before data was simulated by following the multiplicative intensity model as described by Andersen and Gill which does not allow differences in the baseline hazard of the different events. Additionally, as the number of subjects is decreased in the higher ranks of events, the conditional frailty models and the conditional risk set models (PWP-GT and PWP-CP) as well are not able to obtain stable estimates (Lim et al, 2007). Therefore, for a more thorough evaluation of the conditional frailty model, one should draw the simulated data sets from a hazard function that allows for event specific baseline hazards, so that the effect of event dependence would be induced. However, having in mind that within-subject correlation, discontinuous risk intervals and event dependence are really difficult situations to handle especially when one is trying to simulate data, more complex techniques are needed.

Table 5.1: Distribution of the estimated θ in the 1000 simulations

(Scenario A: Heterogeneity)

| | Gap timescale | | Calendar timescale | |
|---|---|---|---|---|
| | Frailty | Conditional frailty | Frailty | Conditional frailty |
| θ ( , ] | Freq (%) | Freq (%) | Freq (%) | Freq (%) |
| 0 – 0.1 | 0 (0) | 406 (40.6) | 0 (0) | 802 (80.2) |
| 0.1 – 0.35 | 15 (1,5) | 64 (6.4) | 7 (0.7) | 19 (1.9) |
| 0.35 – 0.65 | 404 (40.4) | 196 (19.6) | 374 (37.4) | 41 (4.1) |
| 0.65 - 0.75 | 235 (23.5) | 47 (4.7) | 269 (26.9) | 6 (0.6) |
| 0.75 - 1 | 298 (29.8) | 116 (11.6) | 330 (33.3) | 28 (2.8) |
| 1+ | 48 (4.8) | 171 (17.1) | 20 (2) | 104 (10.4) |
| total | 1000 (100) | 1000 (100) | 1000 (100) | 1000 (100) |

(Scenario B: NO Heterogeneity)

| | Gap timescale | | Calendar timescale | |
|---|---|---|---|---|
| | Frailty | Conditional frailty | Frailty | Conditional frailty |
| θ ( , ] | Freq (%) | Freq (%) | Freq (%) | Freq (%) |
| 0 – 0.1 | 893 (89.3) | 867 (86.7) | 924 (92.4) | 898 (89.8) |
| 0.1 – 0.35 | 107 (10.7) | 62 (6.2) | 76 (7.6) | 12 (1.2) |
| 0.35 – 0.65 | 0 (0) | 55 (5.5) | 0 (0) | 60 (6) |
| 0.65 - 0.75 | 0 (0) | 9 (0.9) | 0 (0) | 10 (1) |
| 0.75 - 1 | 0 (0) | 6 (0.6) | 0 (0) | 10 (1) |
| 1+ | 0 (0) | 1 (0.1) | 0 (0) | 10 (1) |
| total | 1000 (100) | 1000 (100) | 1000 (100) | 1000 (100) |

Table 5.2: Simulation results for N=1000, β = 0.8, θ = 0.7

| Timescale | | Mean est. β | Mean standard error | Bias | Coverage rate | Mean est. θ | Rejection rate θ |
|---|---|---|---|---|---|---|---|
| Scenario A: Heterogeneity | | | | | | | |
| Gap | Cox | 0.741 | 0.092 | 0.059 | 0.881 | | |
| | Marginal | 0.741 | 0.101 | 0.059 | 0.913 | | |
| | PWP-GT | 0.671 | 0.093 | 0.129 | 0.706 | | |
| | Frailty | 0.8 | 0.109 | >0.001 | 0.947 | 0.69 | 0.044 |
| | conditional frailty | 0.762 | 0.104 | 0.038 | 0.827 | 0.51 | 0.466 |
| Calendar | Cox | 0.769 | 0.091 | 0.031 | 0.904 | | |
| | AG | 0.769 | 0.106 | 0.031 | 0.936 | | |
| | PWP-CP | 0.667 | 0.093 | 0.132 | 0.694 | | |
| | Frailty | 0.8 | 0.108 | >0.001 | 0.945 | 0.69 | 0.023 |
| | conditional frailty | 0.71 | 0.098 | 0.09 | 0.729 | 0.22 | 0.809 |
| Scenario B: No heterogeneity | | | | | | | |
| Gap | Cox | 0.799 | 0.09 | 0.0015 | 0.957 | | |
| | Marginal | 0.799 | 0.09 | 0.0015 | 0.959 | | |
| | PWP-GT | 0.799 | 0.091 | >0.001 | 0.953 | | |
| | Frailty | 0.8 | 0.09 | >0.001 | 0.955 | 0.02 | 1 |
| | conditional frailty | 0.814 | 0.093 | 0.013 | 0.939 | 0.06 | 0.937 |
| Calendar | Cox | 0.798 | 0.089 | 0.0014 | 0.953 | | |
| | AG | 0.798 | 0.089 | 0.0014 | 0.954 | | |
| | PWP-CP | 0.799 | 0.091 | >0.001 | 0.952 | | |
| | Frailty | 0.799 | 0.089 | >0.001 | 0.953 | 0.01 | 1 |
| | conditional frailty | 0.816 | 0.093 | >0.001 | 0.923 | 0.07 | 0.92 |

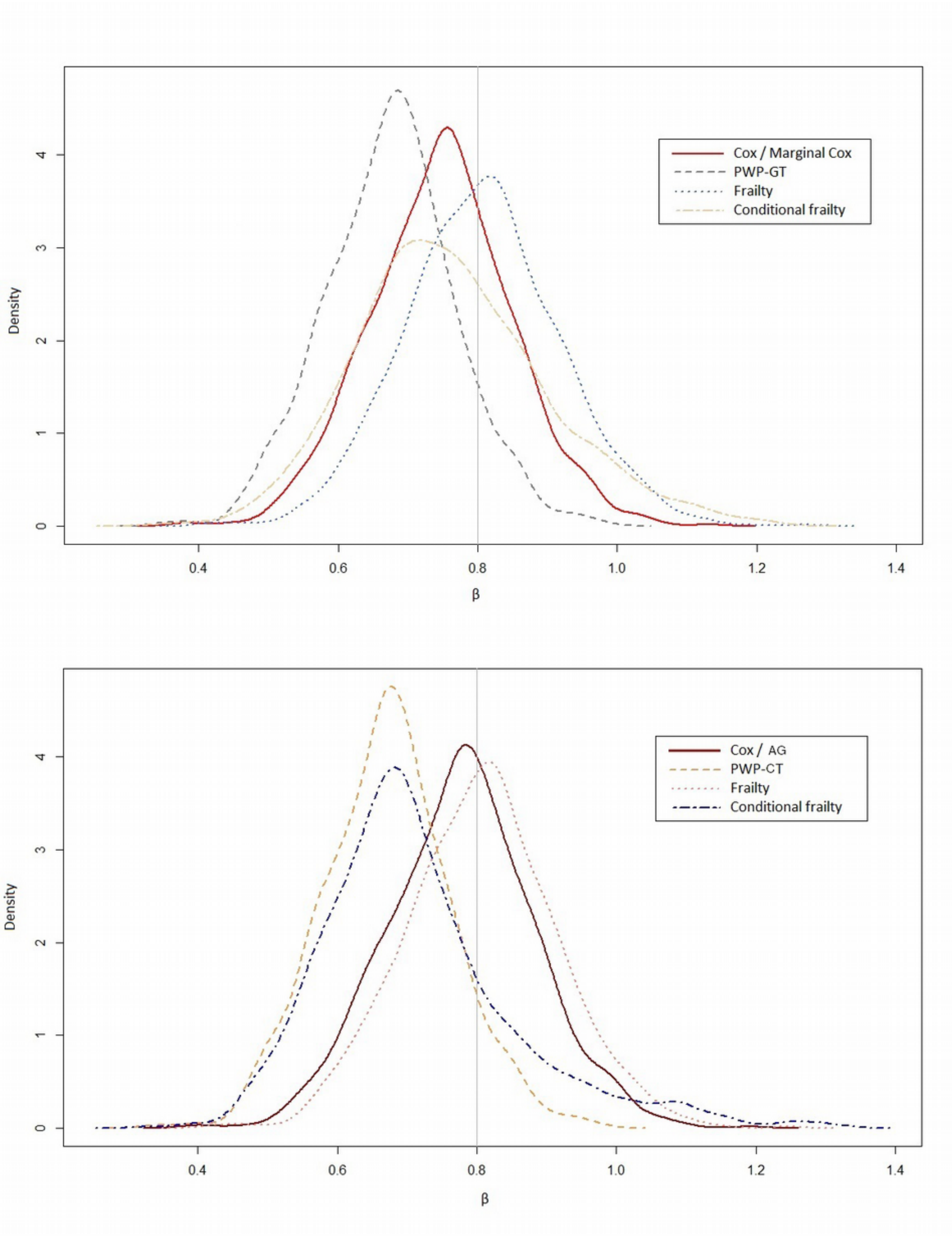


Figure 5.1: Densities of the estimated β (Scenario A: Heterogeneity) Gap and Calendar timescales

# Chapter 6.

# Application in real life data

In this chapter we present the aspects and measurements of our study population for the illustration of all models mentioned in the previous chapters and foremost the Shared Frailty Models as well as the results of the analysis.

## 6.1. Patient Characteristics

For the illustration of the models, we used the data provided by the AMACS. As mentioned before, AMACS within the twenty years of its existence, has monitored over 7000 patients. More specifically, our initial dataset included 7038 HIV-positive patients that had 232037 visits, in total, in the clinical sites spanning from 1 January 1985 to 10 December 2014. On average, every patient had attended the clinical sites 32 times, with the median number of visits being 23.

For our analysis, we focused on patients that had at least one follow-up visit to the clinical site after their first appearance. Therefore, out of the 7038 patients, we excluded 66 patients who had just one visit and never returned to care either because they were loss to follow up or because their visit date was too close to the closure of the dataset, so we were left with 6972 patients who had at least two visits. In addition, 13 more patients were excluded as they were doubts about the accuracy of their entries since appointments were documented after their designated date of death. Furthermore, we excluded 4 patients that had undergone change of sex operation since we had not enough people to form a group with this characteristic and 92 patients that were not adults yet during their first visit in the clinical care since we wanted to limit our study to the patients who were of legal age.

Finally, 31 more patients were excluded since there were missing data in key components for our analysis, such as age, CD4 count or virus load; leaving us with 6832 HIV-positive patients that were included in the analysis.

Table 6.1 presents the descriptive statistics of our study population. The vast majority of our study population is male 5726 (83.8%) whereas there are 1106 (16.2%) females. The study population is representative of the people who live with HIV in Greece since according to KEELPNO, the corresponding percentages are 82% for males and 17.8% for females. As mentioned above, patients who were at least 18 years of age at the initiation of monitoring were included in the analysis, providing a median baseline age of 33.7 years at enrolment, with inter-quartile range equal to (28 – 41) years.

As far as the CD4 cell count is concerned, we used both the original form of the variable and the well documented (Fitzmaurice et al. (2012), Yu et al. (1997), Malaza et al. (2013)) natural logarithm transformation for including this variable in our model. The model with the logarithm included produced a lower value in the AIC, therefore we decided to keep the transformed variable. Additionally, another reason for doing so is justified by the fact that the variable in its raw form may not reflect its actual effect on the outcome. For instance, the effect of a difference of 100 cells between a subject that has 50 cd4 cells/mm$^3$ and a subject that has 150 cd4 cells/mm$^3$ is far greater than it would be if the subjects had 650 and 750 cd4 cells/mm$^3$, respectively. Furthermore, the basic advantage of log transformation is that the interpretation of the transformed variable remains relatively simple. Naturally there were some missing values, so we needed a way to replace them. Therefore, whenever we came across a visit with missing value on the CD4 count, the closest existing measurement was used to replace it. The only restriction was that the replacement measurement should have been registered within a year from the visit. If there was no measurement within a year then no change was made and the missing value remained as missing.

In the case of virus RNA load we followed a similar procedure. First, we applied the analysis on the original form of variable, then the natural logarithm transformation and finally the base 10 logarithmic transformation. Based on the AIC we decided to include the virus load in its base 10 logarithmic transformation. This transformation is also a common practice, as we see in Quinn's work (Quinn et al. (2000)). As done in CD4 count variable, the threshold of a year since each patient's visit was used for replacing missing values. In cases where viral load was below assay's detection limit, viral load was assumed to be equal to the half of the detection limit value.

Another major threat for the condition of an HIV-positive person is the possible coinfections. Apart from turbeculosis which is the most prevalent coinfection with approximately 1/3 of the HIV-infected people being also infected by it (Lancet, 2013), people who live with HIV also report high prevalence of Hepatitis B (HBV) or Hepatitis C (HCV) due to the fact that these viruses share common routes of transmission. According to Alter (2006), an estimated 2-4 million people have chronic HBV coinfection. As of March 2016, an estimated 2.3 million people living with HIV are also infected by Hepatitis C based on the study of Platt et al. (2016). The same study, concluded that people who live with HIV are on average six times more prone to HCV infection than HIV uninfected people. However the total incidence of TB cases among people with HIV in Greece through the years is estimated to 16 by UNAIDS, therefore, we focused on the HBV and HCV coinfections. We assumed that a person has developed one or both of the above-mentioned coinfections if they were tested positive within the first year of observation or had only positive results on the tests throughout the whole observational period.

Our main concern on this thesis is the gaps in care/monitoring. As seen in table 6.2, out of the 6832 subjects of our study 2713 (39.7%) were regularly monitored and treated hence they had no gaps in their care, providing a total observational time of 6,954,489 days until they were censored due to the administrative censoring or due to death. The remaining 4119 (60.3%) subjects had at least one gap in care with the vast majority 4044 (59.1%) having up to 4 gaps in care.

**Table 6.1: Demographical characteristics of study population at initiation of monitoring (N=6832)**

| Patient characteristics at initiation of monitoring | Descriptive statistic[1] |
|---|---|
| Age in years | 33.7 (28-41) |
| Males | 5726 (83.8) |
| Origin | |
| Europe | 5908 (86.5) |
| Africa | 333 (4.9) |
| Asia & Australia | 127 (1.9) |
| Americas | 95 (1.4) |
| Level of education | |
| None/ Primary Education | 175 (2.6) |
| Secondary Education | 1348 (19.7) |
| Higher / Vocational Education | 325 (4.8) |
| Academical Education | 786 (11.5) |
| Possible source of infection | |
| Homosexual / Bisexual Contact | 3479 (50.9) |
| PWID | 606 (8.9) |
| Blood Transfusion | 104 (1.5) |
| Heterosexual Contact | 1704 (24.9) |
| HAARTreatment | 5193 (76) |
| Coinfections[2] | 1033 (15.1) |
| Virus Load (copies/ml) | 21700 (3179-89422) |
| CD4 cell count (cells/μL) | 349 (164-548) |

[1] *for the continuous characteristics we used median (IQR - Inter-quartile Range), whereas for the categorical we used n (%)*

[2] *If a patient developed a coinfection (Hepatitis C or B) either within the 1st year of follow-up or has only positive tests throughout his/her observation period*

Patients with gaps provided either censored observational time in their last entry due to closure of database or death in the case that they returned to care after their last gap or an event time (i.e. missing visit) if they did not return to care after their last gap. The fairly small count of events per subject classified the dependence of event times as event-related dependence (Hougaard, 2000). This dependence can both be negative or positive for the possibility of future events. In Table 6.2, the number of gaps per subject is summarized.

**Table 6.2: Number of gaps during monitoring/treatment**

| Number of gaps | Patients (%) |
|---|---:|
| No gaps | 2713 (39.7) |
| 1 | 2718 (39.7) |
| 2 | 870 (12.7) |
| 3 | 313 (4.6) |
| 4 | 143 (2.1) |
| 5 | 43 (0.6) |
| 6 | 24 (0.3) |
| 7 | 7 (0.1) |
| 8 | 1 (0.01) |

The year with the most enrollments in AMACS is the initiation year (1996) with 930 people starting being monitored and this is equal to 13,6% of all patients ever being monitored. In the following years, the count of newly enrolled patients is rather stable with approximately 280 patients being enrolled every year until 2003. Starting from 2004 a gradual increase in new enrollments is observed till this increase peaks during 2011 and 2012. This is partly justified by the increase of newly infected people among PWID that was mentioned in the introduction of this thesis, as a result of the financial crisis in Greece. In Table 6.3, new enrollments by possible route of infection are presented.

Table 6.3: Count of new enrolments in AMACS per year by possible source of infection[1]

| Year of enrollment | Homosexual / Bisexual contact | IDUs | Blood transfusion | Heterosexual contact |
|---|---|---|---|---|
| | N (%) | N (%) | N (%) | N (%) |
| 1996 | 494 (7.2) | 19 (0.3) | 75 (1.1) | 244 (3.6) |
| 1997 | 188 (2.7) | 10 (0.15) | 4 (0.06) | 107 (1.6) |
| 1998 | 146 (2.1) | 10 (0.15) | 2 (0.03) | 62 (0.9) |
| 1999 | 135 (2) | 11 (0.16) | 6 (0.09) | 100 (1.5) |
| 2000 | 113 (1.6) | 14 (0.2) | 3 (0.04) | 89 (1.3) |
| 2001 | 138 (2) | 10 (0.15) | 1 (0.01) | 88 (1.3) |
| 2002 | 144 (2.1) | 10 (0.15) | 0 (0) | 83 (1.2) |
| 2003 | 107 (1.6) | 7 (0.1) | 2 (0.03) | 93 (1.3) |
| 2004 | 166 (2.4) | 14 (0.2) | 0 (0) | 93 (1.3) |
| 2005 | 219 (3.2) | 5 (0.07) | 2 (0.03) | 105 (1.5) |
| 2006 | 188 (2.7) | 12 (0.18) | 3 (0.04) | 76 (1.1) |
| 2007 | 214 (3.1) | 10 (0.15) | 1 (0.01) | 78 (1.1) |
| 2008 | 204 (3) | 13 (0.19) | 2 (0.03) | 92 (1.3) |
| 2009 | 221 (3.2) | 15 (0.22) | 1 (0.01) | 77 (1.1) |
| 2010 | 255 (3.7) | 24 (0.35) | 0 (0) | 84 (1.2) |
| 2011 | 202 (3) | 117 (1.7) | 1 (0.01) | 92 (1.3) |
| 2012 | 173 (2.5) | 169 (2.5) | 0 (0) | 72 (1.1) |
| 2013 | 140 (2) | 105 (1.5) | 1 (0.01) | 44 (0.6) |
| 2014 | 32 (0.5) | 31 (0.4) | 0 (0) | 25 (0.35) |

[1] Based on the reports of patients

## 6.2. Results

In both timescales, the estimates are obtained in R. COXPH, was the procedure that was used. This procedure practically, fits shared frailty models as a penalized Cox model with penalty function

$$p(w)=(1/\theta)\sum\left[w_i-\exp(w_i)\right]$$

where the $w_i$ are distributed as the logs of iid gamma random variables and their variance is θ.

In many cases, the quantities of interest are the estimations of the regression coefficients and the dependence of the event times is just a nuisance parameter that should be nonetheless assessed, in order to reduce the variance of those estimations. Likewise, in our case, one is more interested in the dependence by itself, however, we need to include explanatory variables in order to reduce the variation owing to the unobserved covariates that comprise the frailty term.

Two simple semiparametric Cox models (Model I and Model VI) were fitted for both timescales. The ordinary Cox model treat each observation as independent in order to estimate the variance of $\hat{\beta}$ . It is obvious that in the case of multivariate analysis, this assumption can be easily violated. Both of these models were fitted in order to compare the results with the ones from the frailty models and assess the effect of the frailty terms.

Consequently, two more cox models were fitted, the marginal model under the gap timescale and the marginal model under the calendar timescale that correspond to the AG model mentioned in chapter 4. These models produced similar estimates of the covariates' coefficients as the simple Cox models. What makes them differ though from the simple case is the assessment of the heterogeneity of subjects through the robust SE estimates. Under the gap timescale, standard errors and robust standard errors remain practically the same, while under the calendar timescale robust standard errors are doubled. That leads to wider confidence intervals, since we use the robust estimates to calculate CI, which means that there is a loss in precision. Nevertheless, what seems to be of more importance between these two cases, is that robust estimates under the gap timescale fail to control for possible heterogeneity between subjects. That could also mean that under the gap timescale there is no significant heterogeneity.

The two conditional models that were fitted for the two timescales (PWP-GT Model III and PWP-CP Model VIII), as we have mentioned before, take into account the natural order of the recurrent events by incorporating event-based baseline hazards. That is the main difference between the two previously fitted models for the gap timescale (Model I and Model II) and the latter one (PWP-GT). Under the gap timescale the estimated HR's remain virtually the same, however there is a big change in the AIC value which is actually the lowest value that we get among the models that were fitted under gap timescale. Similarly, under the calendar timescale we do not observe big changes in the estimated HR's, which are actually even smaller compared to the changes in the gap timescale. In addition, under the calendar timescale, AIC is smaller than the ones from the simple Cox and A-G models, making PWP-CP more appropriate for our dataset.

Next, the two simple frailty models were fitted. Under the gap timescale, the inclusion of the frailty term does not affect the estimated HR's. This is also justified by the almost absent variance θ of the frailty term, which is not significantly different from zero (p-value=0.4 in the corresponding Wald test). In that sense, even the simple Cox model seems to be appropriate to model our data in the gap timescale especially if we consider the complexity of the gamma frailty model.

On the other hand, under the calendar timescale, the inclusion of the frailty term seems to account for the variability between subjects. The estimated covariates' coefficients differed from those obtained from the previous models without the frailty term while the variance of the θ is estimated to be 0.6051, corresponding to 2393 degrees of freedom. In addition, the standard errors of the coefficients reported by the frailty model were closer to the ones that were reported in the marginal model with the robust correction. The approximate Wald test for the frailty is a chi-square test with a value 3618.2 resulting to a p-value<0.001 implying that controlling for the within-subject correlation makes sense. This model, is actually an AG model, as described earlier, with a gamma frailty accounting for the correlation of the event times. Therefore, it treats event times, in our case the gaps in care, as ordered outcomes while data is set up in the counting process form. In particular, by exponentiating the standard deviation of 0.7779 we get 2.177, meaning that patients with a frailty of one standard deviation above the mean, are 117.7% more frail (have greater risk) to commit a gap than a patient with frailty equal to mean frailty and the same observed covariate values. A likelihood ratio test (LR test) was performed for the frailty term. This test is twice the difference between the log partial likelihood with the frailty terms integrated out (indicated as I-likelihood on the R printout) and the likelihood of a no-frailty model (ordinary Cox Model). The test is equal to 2 (49821.41-49456.7)=729.42, which is chi-square with 1 df. The corresponding p-value is <0.0001 implying that the frailty term is significantly related to the time to recurrence of a gap.

The LR test given on the printout of the R-procedure for the frailty model is a test for the model as a whole and it is the difference between the unpenalized log-partial likelihood at the final iteration and at the initial values of the full set of parameters. One can easily see that with the values at the initial step set β=0 and u=0 the likelihood reduces to the one of the no-frailty model (Cox) when β values are also equal to 0. Gray (1992) showed that the actual distribution of this test is a weighted sum of chi squares. Therefore the test for the regression parameters equal to zero is rejected with chi-square 6810 for 2404 df and p-value<0.0001.

The conditional frailty models were also fitted. These models allow for the possibility that both heterogeneity of subjects and dependence of the events play a significant role in the subjects risk for a particular (re)occurrence since it incorporates a random effect and an event-specific baseline hazard as well. Once again the natural order of the events is taken into account, this time by incorporating event-based stratification. We notice that under the gap timescale, the effect estimates of the covariates, their standard errors and the corresponding CI's are virtually unchanged compared to the simple shared frailty model. The variance θ, for this model, is practically equal to 0, meaning that the inclusion of the frailty term does not account for any unobserved variability. On the other hand, in Model X (Conditional shared frailty model) we observe that the variance θ of frailty term is 0.7059, which is a little larger than the corresponding one derived from the simple frailty model under the calendar timescale (Model IX). Once again, by exponentiating the standard deviation of 0.8402 (the square root of variance θ=0.7059) we get 2.31, meaning that patients with a frailty of one standard deviation above the mean, are 131% more frail to commit a gap than a patient with frailty equal to mean frailty and the same observed covariate values.

Accounting for the dependence of the events by stratification does not seem to alter drastically the results of the simple frailty model and the conditional frailty model in the two timescales respectively, apart from the slight increase of the variance θ in the calendar timescale, as the estimated HR's and SE's of the conditional frailty models are essentially the same as those of the simple frailty fits when compared under the same timescale. However the conditional frailty model (Model X) produces the lowest AIC value compared to all other models in both timescales making it the most appropriate model for our dataset.

In figure 6.1 the Martingale residuals for each model are presented. As mentioned in chapter 3, the Martingale residuals are plotted against the variable representing age at initiation of monitoring, which is a time-constant variable. The advantages of plotting them against such variable are also mentioned in chapter 4. However, since it is just a visual evaluation, we should interpret it as a simple indication of the goodness of fit. These residuals can be viewed as the difference between the observed number of fails for each subject and the expected number of fails, based on the fitted models. The two models that perform better according to the AIC (i.e., model VIII and model X) perform better also when we evaluate their fit through also seem to be performing better with regard to the residuals as well. Both PWP-CP and conditional frailty model under the calendar timescale have their residuals centered about zero. However, the residuals of the PWP-CP present higher values hence larger differences between the observed and the expected number of fails. Therefore, based on the Martingale residuals, the conditional frailty model is the most appropriate for our dataset.

Based on all the above, we conclude that the best model to describe our data is the conditional frailty model under the calendar timescale thus we will extend into interpreting its estimated coefficients.

As we mentioned before, there are two sources of correlation for the event times of recurrent events. One is the heterogeneity of the individuals which implies a correlation of event times on an individual level and the other is the event correlation itself, where the occurrence of an event affects the possibility of a future event. The first source of correlation is assessed by the frailty term whereas the second source is assessed by the event based stratification.

Under the gap timescale, we are trying to assess the effect of a recurrent gap in care on the recurrent event rate of the subsequent gap. The fits of the models indicate that heterogeneity of the subjects plays no significant role in the evolution of the event rate, which is measured by the interval between two subsequent gaps. This is justified by the almost zero estimated variance of the frailty term, meaning that frailty models under the gap timescale do not identify subjects more frail to commit gaps in care. On the other hand, the results of the PWP-GT model, point to the existence of event correlation as it allows for event based baseline hazards.

Under the calendar timescale, we are trying to assess the effect of a recurrent gap on the event rate as a function of time since the initiation of monitoring. The conditional frailty model indicates that in our data both sources of correlation are present. The frailty term indicates that some subjects are more or less prone to commit a gap and this is a function of time since initiation of monitoring, meaning that the longer someone stays under observation the more his behavior against committing or not a gap is affected. Event correlation is also apparent in our data, indicating that the natural order of the gaps plays also a significant role in the event rate since initiation of monitoring.

The factor with the most impact in treatment retention and subsequently in the recurrent event rate seems to be the one indicating if patient receives HAART or not. People who do not receive HAART tend to commit gaps more often compared to people who receive treatment. Specifically, people who receive HAART have a 55% reduced hazard of committing a gap compared to those who do not receive HAART holding all other covariates stable at a given time t. Women are classified as more punctual than men since according to the model results as the rate of recurrence among them is approximately 0.92 times the rate among men, holding all other covariates stable at any given time t. The origin of the subject also seems to be important for the outcome since people coming from Europe are the less frail people among the study population. People coming from Africa could be identified as the most endangered group of origin to commit recurrent gaps as they face an increased hazard rate of 1,6 times the hazard Europeans face, holding all other covariates stable.

One would think that having a coinfection would make someone more punctual to his appointments, however results point to the opposite direction. Having a coinfection increases the hazard rate of recurrent event by almost 14% compared to the hazard of the subjects that do not have a coinfection. One possible reason for that, could be the fact that almost half of the people with a coinfection (48%), are people who also inject drugs (PWID). Those people are also more frail compared to those infected by other routes when compared to the reference group that reported heterosexual contact as the possible means of transmission. More specifically they present a 16% increase in their hazard rate compared to people who indicate heterosexual contact as the most possible means of infection, holding all other covariates stable at time t.

Age at initiation of monitoring also affects the recurrent event rate meaning that the older someone starts to be monitored the less frail is to commit recurrent gaps. Although there is an almost 2% decrease in the hazard rate for an increase of a year in the age of the subject, often a one-year change is not of clinical importance. For instance, let us see what happens in a five year change in age. For every five year increase in the age of the subject an increase of 7% (0.985^5) in the hazard rate is observed and it is independent of the age at which the increase is calculated.

Table 6.4: Results of Cox, Marginal and PWP-GT models under the Gap Timescale

| Covariates | Cox Model I | | | Marginal Cox Model II | | | PWP-GT Model III | | |
|---|---|---|---|---|---|---|---|---|---|
| | HR | SE | 95% CI | HR | robust SE | 95% CI* | HR | robust SE | 95% CI* |
| Age (years) | 0.988 | 0.001 | (0.985,0.99) | 0.988 | 0.001 | (0.985,0.99) | 0.990 | 0.001 | (0.987,0.992) |
| Sex | 0.936 | 0.039 | (0.865,1.013) | 0.936 | 0.041 | (0.862,1.017) | 0.947 | 0.039 | (0.876,1.023) |
| Origin | | | | | | | | | |
| Europe | 1 | | | 1 | | | 1 | | |
| Africa | 1.32 | 0.079 | (1.181,1.475) | 1.32 | 0.074 | (1.189,1.466) | 1.339 | 0.071 | (1.213,1.479) |
| Asia & Australia | 1.253 | 0.128 | (1.045,1.503) | 1.253 | 0.123 | (1.051,1.493) | 1.288 | 0.118 | (1.092,1.52) |
| Americas | 1.03 | 0.141 | (0.813,1.306) | 1.03 | 0.139 | (0.815,1.302) | 1.081 | 0.139 | (0.864,1.354) |
| Not specified | 1.088 | 0.065 | (0.973,1.216) | 1.088 | 0.070 | (0.966,1.225) | 1.115 | 0.067 | (0.998,1.246) |
| Education | | | | | | | | | |
| None / Primary education | 1 | | | 1 | | | 1 | | |
| Secondary Education | 0.855 | 0.075 | (0.73,1.002) | 0.855 | 0.076 | (0.728,1.004) | 0.863 | 0.072 | (0.742,1.004) |
| Higher / Vocational Educati | 0.876 | 0.093 | (0.725,1.059) | 0.876 | 0.095 | (0.723,1.063) | 0.893 | 0.09 | (0.746,1.07) |
| Academical Education | 0.903 | 0.083 | (0.765,1.066) | 0.903 | 0.086 | (0.761,1.072) | 0.908 | 0.081 | (0.774,1.066) |
| Not specified | 1.027 | 0.085 | (0.884,1.194) | 1.027 | 0.086 | (0.883,1.195) | 1.041 | 0.081 | (0.902,1.2) |
| Possible source of Infection | | | | | | | | | |
| Heterosexual contact | 1 | | | 1 | | | 1 | | |
| Homosexual/Bisexual conta | 0.825 | 0.033 | (0.766,0.889) | 0.825 | 0.035 | (0.763,0.893) | 0.840 | 0.033 | (0.781,0.904) |
| PWID | 1.078 | 0.075 | (0.948,1.225) | 1.078 | 0.077 | (0.945,1.229) | 1.112 | 0.074 | (0.983,1.257) |
| Blood transfusion | 0.557 | 0.074 | (0.442,0.701) | 0.557 | 0.094 | (0.418,0.741) | 0.578 | 0.092 | (0.441,0.758) |
| Not specified | 0.972 | 0.045 | (0.892,1.06) | 0.972 | 0.046 | (0.889,1.063) | 0.990 | 0.044 | (0.911,1.076) |
| Haart | 0.581 | 0.019 | (0.546,0.619) | 0.581 | 0.018 | (0.548,0.617) | 0.557 | 0.017 | (0.526,0.59) |
| Coinfections | 1.104 | 0.050 | (1.014,1.202) | 1.104 | 0.057 | (1.003,1.216) | 1.090 | 0.052 | (0.997,1.191) |
| Virus load (copies/ml) | 0.963 | 0.010 | (0.945,0.982) | 0.963 | 0.010 | (0.944,0.983) | 0.987 | 0.01 | (0.968,1.006) |
| CD4 count (cells/µL) | 0.997 | 0.013 | (0.973,1.022) | 0.997 | 0.013 | (0.973,1.022) | 0.999 | 0.012 | (0.976,1.023) |
| logLikelihood | -54388.850 | | | -54388.85 | | | -47883.11 | | |
| AIC | 108813.7 | | | 108813.7 | | | 95802.22 | | |

* Confidence intervals based on the Robust SE

Table 6.5 : Results of Frailty and Conditional Frailty models under the Gap timescale

| Covariates | Frailty Model IV | | | Conditional frailty Model V | | |
|---|---|---|---|---|---|---|
| | HR | SE | 95% CI* | HR | SE | 95% CI* |
| Age (years) | 0.988 | 0.002 | (0.986,0.991) | 0.990 | 0.002 | (0.988,0.993) |
| Sex | 0.942 | 0.04 | (0.87,1.02) | 0.952 | 0.040 | (0.879,1.03) |
| Origin | | | | | | |
| Europe | 1 | | | 1 | | |
| Africa | 1.33 | 0.08 | (1.19,1.486) | 1.348 | 0.081 | (1.207,1.506) |
| Asia & Australia | 1.264 | 0.129 | (1.054,1.517) | 1.299 | 0.132 | (1.083,1.558) |
| Americas | 1.038 | 0.142 | (0.818,1.316) | 1.089 | 0.149 | (0.859,1.381) |
| Not specified | 1.094 | 0.066 | (0.979,1.223) | 1.121 | 0.067 | (1.003,1.253) |
| Education | | | | | | |
| None / Primary education | 1 | | | 1 | | |
| Secondary Education | 0.857 | 0.075 | (0.731,1.004) | 0.865 | 0.076 | (0.738,1.014) |
| Higher / Vocational Education | 0.877 | 0.094 | (0.726,1.061) | 0.894 | 0.095 | (0.74,1.081) |
| Academical Education | 0.902 | 0.083 | (0.763,1.065) | 0.907 | 0.084 | (0.768,1.071) |
| Not specified | 1.027 | 0.085 | (0.883,1.194) | 1.039 | 0.086 | (0.894,1.208) |
| Possible source of Infection | | | | | | |
| Heterosexual contact | 1 | | | 1 | | |
| Homosexual/Bisexual contact | 0.825 | 0.033 | (0.766,0.889) | 0.840 | 0.034 | (0.78,0.906) |
| PWID | 1.085 | 0.076 | (0.954,1.234) | 1.118 | 0.078 | (0.984,1.271) |
| Blood transfusion | 0.561 | 0.074 | (0.445,0.706) | 0.583 | 0.077 | (0.463,0.734) |
| Not specified | 0.974 | 0.045 | (0.893,1.062) | 0.991 | 0.046 | (0.909,1.081) |
| Haart | 0.58 | 0.019 | (0.545,0.618) | 0.557 | 0.019 | (0.523,0.594) |
| Coinfections | 1.103 | 0.05 | (1.013,1.201) | 1.089 | 0.049 | (1.001,1.185) |
| Virus load (copies/ml) | 0.964 | 0.01 | (0.946,0.983) | 0.987 | 0.010 | (0.968,1.007) |
| CD4 count (cells/μL) | 0.995 | 0.013 | (0.971,1.02) | 0.997 | 0.013 | (0.973,1.022) |
| Variance of θ | 0.003 | | | <0.001 | | |
| logLikelihood | -54541.72 | | | -48058.06 | | |
| AIC | 109085.4 | | | 96118.11 | | |

*Confidence intervals were computed under the assumption that the variance of θ was fixed

Table 6.6: Results of Cox, Marginal and PWP-CP models under the Calendar Timescale

| Covariates | Cox Model VI | | | A-G (Marginal Cox) Model VII | | | PWP-CP Model VIII | | |
|---|---|---|---|---|---|---|---|---|---|
| | HR | SE | 95% CI | HR | robust SE | 95% CI* | HR | robust SE | 95% CI* |
| Age (years) | 0.986 | 0.001 | (0.983,0.988) | 0.986 | 0.002 | (0.982,0.989) | 0.989 | 0.002 | (0.985,0.992) |
| Sex | 0.904 | 0.038 | (0.835,0.978) | 0.904 | 0.049 | (0.816,1.001) | 0.914 | 0.046 | (0.832,1.004) |
| Origin | | | | | | | | | |
| Europe | 1 | | | 1 | | | 1 | | |
| Africa | 1.512 | 0.090 | (1.354,1.688) | 1.512 | 0.112 | (1.32,1.732) | 1.512 | 0.103 | (1.333,1.714) |
| Asia & Australia | 1.386 | 0.141 | (1.156,1.663) | 1.386 | 0.179 | (1.107,1.736) | 1.436 | 0.17 | (1.166,1.769) |
| Americas | 1.226 | 0.168 | (0.967,1.555) | 1.226 | 0.203 | (0.927,1.623) | 1.318 | 0.206 | (1.009,1.721) |
| Not specified | 1.06 | 0.064 | (0.947,1.186) | 1.06 | 0.088 | (0.911,1.233) | 1.070 | 0.079 | (0.935,1.225) |
| Education | | | | | | | | | |
| None / Primary education | 1 | | | 1 | | | 1 | | |
| Secondary Education | 0.82 | 0.071 | (0.7,0.96) | 0.82 | 0.092 | (0.672,1) | 0.848 | 0.084 | (0.71,1.013) |
| Higher / Vocational Education | 0.867 | 0.092 | (0.718,1.048) | 0.867 | 0.117 | (0.686,1.096) | 0.915 | 0.109 | (0.741,1.128) |
| Academical Education | 0.887 | 0.082 | (0.751,1.047) | 0.887 | 0.107 | (0.717,1.096) | 0.915 | 0.097 | (0.758,1.105) |
| Not specified | 1.04 | 0.086 | (0.894,1.209) | 1.04 | 0.111 | (0.86,1.257) | 1.081 | 0.103 | (0.912,1.282) |
| Possible source of Infection | | | | | | | | | |
| Heterosexual contact | 1 | | | 1 | | | 1 | | |
| Homosexual/Bisexual contact | 0.812 | 0.032 | (0.754,0.875) | 0.812 | 0.042 | (0.737,0.894) | 0.835 | 0.039 | (0.765,0.911) |
| PWID | 1.319 | 0.093 | (1.158,1.501) | 1.319 | 0.121 | (1.117,1.557) | 1.350 | 0.111 | (1.162,1.568) |
| Blood transfusion | 0.468 | 0.062 | (0.372,0.59) | 0.468 | 0.092 | (0.338,0.649) | 0.515 | 0.094 | (0.38,0.699) |
| Not specified | 0.95 | 0.044 | (0.871,1.037) | 0.95 | 0.057 | (0.851,1.062) | 0.985 | 0.053 | (0.891,1.089) |
| Haart | 0.474 | 0.015 | (0.445,0.504) | 0.474 | 0.018 | (0.439,0.51) | 0.455 | 0.017 | (0.424,0.488) |
| Coinfections | 1.154 | 0.052 | (1.06,1.256) | 1.154 | 0.074 | (1.025,1.299) | 1.129 | 0.063 | (1.017,1.253) |
| Virus load (copies/ml) | 0.963 | 0.010 | (0.944,0.982) | 0.963 | 0.012 | (0.939,0.987) | 0.998 | 0.012 | (0.975,1.022) |
| CD4 count (cells/μL) | 1.007 | 0.013 | (0.982,1.032) | 1.007 | 0.015 | (0.977,1.037) | 1.010 | 0.014 | (0.982,1.038) |
| logLikelihood | -49821.41 | | | -49821.41 | | | -43328.95 | | |
| AIC | 99678.82 | | | 99678.82 | | | 86693.89 | | |

* Confidence intervals based on the Robust SE

Table 6.7 : Results of Frailty and Conditional Frailty models under the Calendar timescale

| Covariates | Frailty Model IX | | | Frailty stratified by order of events Model X | | |
|---|---|---|---|---|---|---|
| | HR | SE | 95% CI* | HR | SE | 95% CI* |
| Age (years) | 0.986 | 0.002 | (0.983,0.989) | 0.985 | 0.002 | (0.982,0.989) |
| Sex | 0.925 | 0.054 | (0.83,1.03) | 0.918 | 0.056 | (0.821,1.027) |
| Origin | | | | | | |
| Europe | 1 | | | 1 | | |
| Africa | 1.624 | 0.141 | (1.388,1.901) | 1.616 | 0.147 | (1.372,1.904) |
| Asia & Australia | 1.507 | 0.216 | (1.177,1.93) | 1.499 | 0.224 | (1.16,1.938) |
| Americas | 1.253 | 0.227 | (0.925,1.698) | 1.236 | 0.231 | (0.904,1.689) |
| Not specified | 1.092 | 0.094 | (0.935,1.276) | 1.077 | 0.096 | (0.917,1.265) |
| Education | | | | | | |
| None / Primary education | 1 | | | 1 | | |
| Secondary Education | 0.793 | 0.098 | (0.638,0.986) | 0.786 | 0.102 | (0.627,0.986) |
| Higher / Vocational Education | 0.841 | 0.125 | (0.651,1.086) | 0.835 | 0.130 | (0.64,1.09) |
| Academical Education | 0.879 | 0.114 | (0.7,1.103) | 0.863 | 0.117 | (0.682,1.093) |
| Not specified | 1.053 | 0.124 | (0.855,1.296) | 1.050 | 0.129 | (0.846,1.303) |
| Possible source of Infection | | | | | | |
| Heterosexual contact | 1 | | | 1 | | |
| Homosexual/Bisexual contact | 0.783 | 0.042 | (0.708,0.866) | 0.775 | 0.043 | (0.698,0.86) |
| PWID | 1.227 | 0.117 | (1.034,1.456) | 1.160 | 0.115 | (0.972,1.385) |
| Blood transfusion | 0.422 | 0.074 | (0.313,0.568) | 0.423 | 0.078 | (0.311,0.575) |
| Not specified | 0.968 | 0.062 | (0.86,1.09) | 0.974 | 0.065 | (0.862,1.101) |
| Haart | 0.439 | 0.019 | (0.404,0.477) | 0.446 | 0.020 | (0.409,0.486) |
| Coinfections | 1.146 | 0.071 | (1.022,1.285) | 1.134 | 0.073 | (1.007,1.277) |
| Virus load (copies/ml) | 0.961 | 0.012 | (0.938,0.984) | 0.950 | 0.013 | (0.926,0.975) |
| CD4 count (cells/μL) | 1.013 | 0.017 | (0.982,1.046) | 1.018 | 0.017 | (0.985,1.051) |
| Variance of θ | 0.6051 | | | 0.7059 | | |
| logLikelihood | -47004.51 | | | 40400.14 | | |
| AIC | 94010.18 | | | 80801.38 | | |

*Confidence intervals were computed under the assumption that the variance of θ was fixed

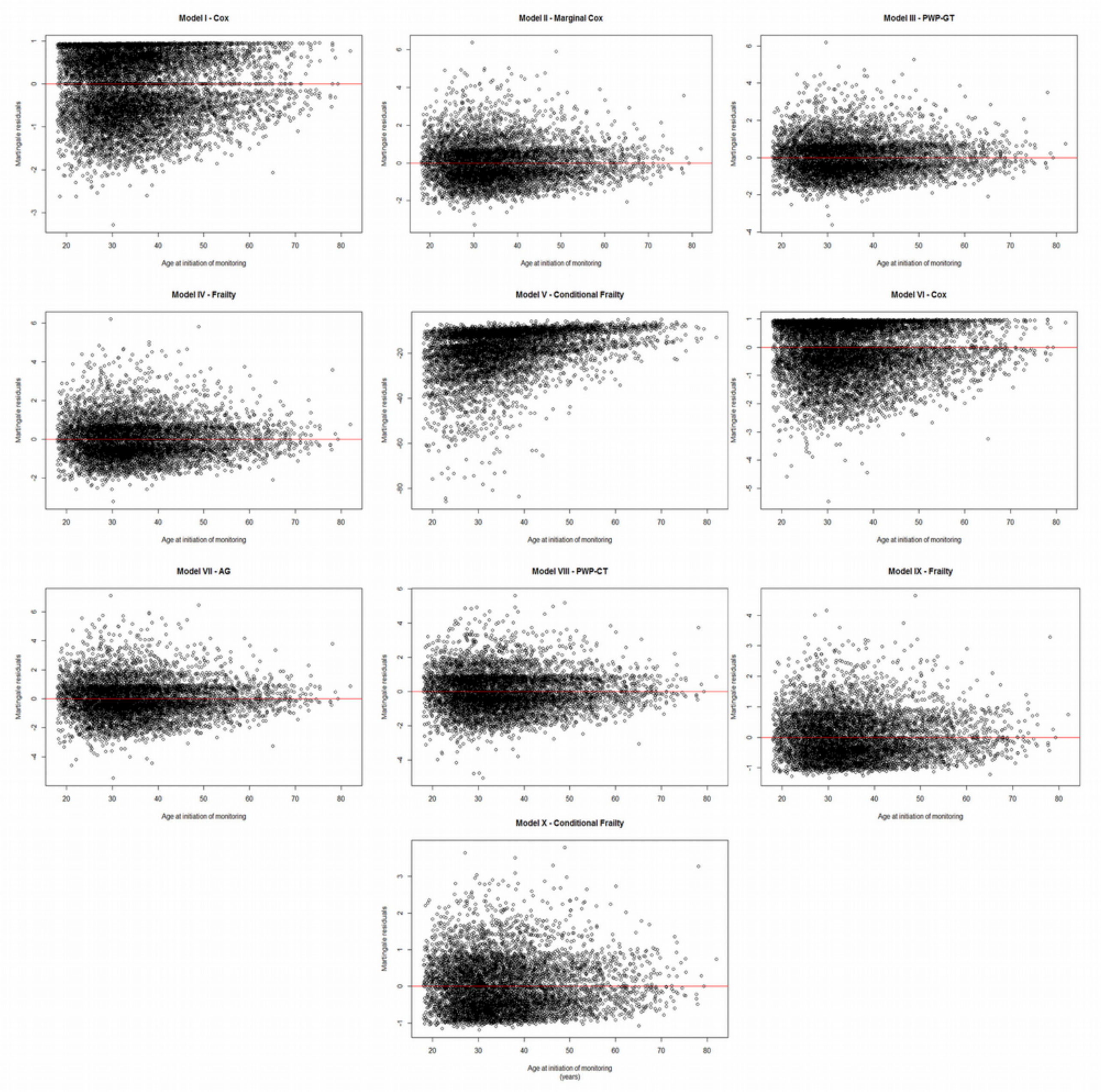
Table 6.8 : Bootstrapping Results of Frailty and Conditional Frailty models under the Gap and Calendar timescale

| Covariates | Frailty Model IV | | Conditional frailty Model V | | Frailty Model IX | | Conditional frailty Model X | |
|---|---|---|---|---|---|---|---|---|
| | HR | SE | HR | SE | HR | SE | HR | SE |
| Age (years) | 0.988 | 0.003 | 0.990 | 0.003 | 0.959 | 0.003 | 0.985 | 0.003 |
| Sex | 0.916 | 0.115 | 0.969 | 0.116 | 0.942 | 0.113 | 0.934 | 0.112 |
| Origin | | | | | | | | |
| Europe | 1 | | 1 | | 1 | | 1 | |
| Africa | 1.342 | 0.237 | 1.363 | 0.241 | 1.62 | 0.286 | 1.625 | 0.287 |
| Asia & Australia | 1.24 | 0.406 | 1.272 | 0.416 | 1.455 | 0.467 | 1.445 | 0.473 |
| Americas | 1.022 | 0.489 | 1.066 | 0.511 | 1.224 | 0.586 | 1.215 | 0.582 |
| Not specified | 1.096 | 0.195 | 1.115 | 0.199 | 1.094 | 0.195 | 1.067 | 0.190 |
| Education | | | | | | | | |
| None / Primary education | 1 | | 1 | | 1 | | 1 | |
| Secondary Education | 0.863 | 0.243 | 0.865 | 0.244 | 0.797 | 0.225 | 0.798 | 0.224 |
| Higher / Vocational Education | 0.887 | 0.31 | 0.892 | 0.312 | 0.843 | 0.295 | 0.085 | 0.296 |
| Academical Education | 0.936 | 0.278 | 0.931 | 0.277 | 0.913 | 0.271 | 0.905 | 0.269 |
| Not specified | 1.048 | 0.277 | 1.048 | 0.277 | 0.931 | 0.246 | 0.928 | 0.245 |
| Possible source of Infection | | | | | | | | |
| Heterosexual contact | 1 | | 1 | | 1 | | 1 | |
| Homosexual/Bisexual contact | 0.834 | 0.094 | 0.843 | 0.095 | 0.791 | 0.089 | 0.783 | 0.088 |
| PWID | 1.08 | 0.226 | 1.108 | 0.231 | 1.201 | 0.251 | 1.155 | 0.241 |
| Blood transfusion | 0.519 | 0.252 | 0.539 | 0.262 | 0.399 | 0.194 | 0.398 | 0.193 |
| Not specified | 0.99 | 0.131 | 1.008 | 0.013 | 0.989 | 0.131 | 0.996 | 0.131 |
| Haart | 0.579 | 0.053 | 0.557 | 0.051 | 0.444 | 0.041 | 0.450 | 0.041 |
| Coinfections | 1.106 | 0.146 | 1.101 | 0.142 | 0.162 | 0.149 | 1.148 | 0.148 |
| Virus load (copies/ml) | 0.972 | 0.025 | 0.994 | 0.026 | 0.969 | 0.259 | 0.962 | 0.025 |
| CD4 count (cells/µL) | 1 | 0.034 | 1.002 | 0.034 | 1.018 | 0.035 | 1.023 | 0.035 |
| Variance of θ | 0.0222 | | <0.001 | | 0.5694 | | 0.6344 | |

As we notice in the tables above, the standard error for the coefficients and the corresponding CI's were computed under the assumption that the variance θ was fixed. This is a smoothing spline for the computation of the standard errors and while this can be a correct assumption for some models, in the frailty framework it is not (Therneau & Grambsch, 2000). In order to check the accuracy of those results, we performed a bootstrap methodology by creating 100 subsets of our data and computing once again these SE's and CI's. The results of the bootstrapping are presented in table 6.8.

One can see that the bootstrapping did not seriously affected effect of the coefficients' estimates compared to those derived from the main analyses (i.e. prior to performing the bootstrap). However, in all cases and timescales we observe that the standard errors have doubled and in some cases have increased even more. This is normal, as the bootstrapping involves re-sampling of the original dataset and it's time the sample is smaller than the complete dataset. As far as the variance of the frailty term is concerned no change in the significance is observed under the gap timescale. However, under the calendar timescale, the variance has been reduced from 0.706 to 0.63 and from 0.6 to 0.58 in the conditional frailty and the simple frailty models respectively.

Figure 6.1 : Martingale residuals for fitted models



*For all models apart Model I and Model VI (Simple Cox), summed Martingale residuals are presented*

## 6.3. Discussion

This thesis gave us the opportunity of getting familiar with very large datasets. The most crucial part when analysing recurrent time to event data is getting data in the required form for the analysis. Different models require different setting of the data and one should be aware of the complications of a dataset of that size and also be able to resolve any issues that arise. Therefore it is no surprise when many authors rely to the illustration of the data layout when trying to explain the differences between all available models for the analysis of recurrent events.

In our analysis apart from using different models for the assessment of gaps in care/treatment, we used two different timescales as well. Additionally, we implemented discontinuous risk intervals, where the subject is not at risk of another failure while a previous one is ongoing. Both timescales, gap and calendar, are set so subjects that are not under risk are excluded from the risk set. This was achieved by taking into account the risk free period of each subject, meaning that if someone is experiencing a gap in treatment / monitoring is not considered under risk for a subsequent gap. This subject re-enters the risk set if only he or she returns to treatment / monitoring after this ongoing gap.

It is important to understand that both timescales we used, model the same risk period for each patient, as mentioned before. The difference between these two timescales is in the definition of the origin of time and therefore in how the risk sets for each model are comprised. Gap timescale measures the time between two subsequent events and every time the clock resets at time zero, meaning that for one person that has for instance 3 gaps in care, his corresponding contribution to the risk set of time 0 is three survival periods, one for each event. This implies that the subject is considered to be at risk for all three events simultaneously. Calendar timescale also measures the time between two subsequent events but this time the clock is not reset at time 0 but instead time continues to run along the actual time since initiation of monitoring. Therefore the same person would contribute just one survival time in the risk set of time 0, which intuitively, seems more rational.

In addition, we should clarify that ignoring the natural order in which recurrent events occur can produce misleading conclusions as not all the available information is taken into account. Hence gap models that do not stratify the analysis by the natural order of the events may lead to inefficient inferences. Moreover, by not stratifying and therefore by assuming a common baseline hazard for each event implies that recurrent events are unaffected by previous events which can also lead to inefficient estimates when event dependence is apparent. However, a problem that occurs by event based stratification is that the number of subjects is limiting while the number of recurrence is increasing, leading to unstable coefficient estimates. In our case though, where the highest rank of occurrence is limited to the number 8, this problem did not seem to have affected the analysis, as in cases where the frequency of the recurrences is small we may assume that the risk of occurrence vary substantially between recurrences (Lim et al., 2007)

Therefore, event based stratification is used in the conditional risk set model in order to account for the natural order of events. By doing so, we adjust for the fact that an occurrence may have been affected by the previous events. Hence, even though the PWP-GT uses the gap configuration of the data, the order of events is induced to the model contrary to the other gap models, so that the hazard functions at time t for the k-th recurrence are conditional on the k-1 previous occurrences. Once again, the main difference between PWP-GT and PWP-CP becomes more clear by the computational view of the construction of the risk-sets. If the researcher suspects that survival times are also correlated within a subject then the addition of the frailty approach should also be implemented as ignoring the frailty when it is present can lead to underestimated covariate effects (Henderson & Oman, 1999). This approach was carried out by the implementation of the conditional frailty models. The frailty term is incorporated for each subject, acting multiplicatively on the baseline hazard (Wienke, 2010) to adjust for unobserved risk factors, leading to variations in the baseline hazard from subject to subject (Lim et al., 2007).

Another major but also subtle difference though, occurs during the interpretation of the estimated coefficients. Duchateau and Janssen (2003) explain that the difference between the two timescales is that the gap timescale investigates the effect of a recurrent event on the event rate of the subsequent event, while calendar timescale investigates the full course of evolution of the recurrent event rate since initiation of monitoring, while using parametric and semiparametric frailty models. In other words, under the gap timescale after an event a new survival time begins which is independent of the previous survival times, while under the calendar timescale, the risk of a recurrence at one point in time depends on the entire path of the attendance behavior of a subject starting at the beginning of the observation window. Hence, under the gap timescale we investigate how the average length of the monitoring/treatment period depends on the given covariates (Clement & Strawderman, 2009) while under the calendar timescale we investigate how covariates affect the full course of the recurrent event rate since initiation of monitoring/treatment.

However, we should mention that this interpretation is based on the fact that coefficient estimates are conditional on the unobserved frailty. Thus the coefficients should be interpreted as the effect of a risk factor on a typical subject, so if we have a boolean covariate, then its effect is interpreted as the relative risk between two typical individuals of the two different categories of the covariate.

In our analysis we used a gamma distribution for the frailty term and even though this is the standard assumption about the frailty, there are many disadvantages especially in regression models. A basic disadvantage of the gamma frailty model is that the within subject correlation is modeled by a single parameter which is the variance of the frailty distribution. Other distributions may be studied to avoid these shortcomings such as positive stable distributions (Hougaard, 2000). By making the assumption of gamma distributed frailty, we relied on the EM algorithm, which as mentioned before, yields a discrete estimator and does not allow direct estimation of the hazard function. PPL is an alternative procedure that is widely used for assessing this problem. Rondeau (2003) proposed a procedure that penalizes the full likelihood instead of the frailties, as proposed by Therneau and Grambsch (2000).

In our example, data collection is an ongoing process that goes back twenty years, so it is proper and very useful to consider covariates that change over time. The covariates that change over time are called time-depended covariates. Time-dependent covariates are divided into two categories, the internal and the external covariates. Internal covariates are the ones that relate to the subject under observation itself, such as weight or CD4 cell count and the external variables that do not depend on the physical observation of the subject.

In addition, when analysing time to event data in a longitudinal study, covariates which are recorded at the time of study entry are less likely to be influential than the more recent values of those covariates. For that reason, we included the available internal time-varying covariates (CD4 log and virus load) in our analysis.

Another term that should not be confused with the time-depended covariates is the time-depended coefficients. In order to distinguish the difference between these two extensions of the Cox model let us see how the hazard function becomes in the simplest forms. In the presence of time-depended covariates the hazard function is given by:

$$\lambda\left(t\right)=\lambda_0\left(t\right)\exp\left(\beta Z\left(t\right)\right)$$

While in the presence of time-depended coefficients the hazard function becomes:

$$\lambda(t)=\lambda_0(t)\exp(\beta(t)Z)$$

The latter expression implies that the proportional hazards assumption no longer stands since the basic assumption of proportional hazards is that the coefficient is time invariant (β(t)=c). As we mentioned before, the first expression suggests that the proportional hazards assumption does not stand as well. However, proportional hazards assumption can be implied since the hazard ratio does not depend on time.

In our application, we made the assumption that the coefficients of the variables were constant over time. However, this may not be rational especially in studies where the follow-up period is that long. Murphy and Sen (1990) included time-depended coefficients in a Cox-type model by using a sieve estimation procedure (Grenander, 1981) to estimate the coefficient. Yu et al. (2013) estimated time-dependent coefficients by implementing penalized spline methods.

As we have mentioned before, Duchateau & Janssen (2002) studied different timescales in the framework of parametric frailty models taking into account the duration of the event. Both their results and ours are pointing to the same direction of using both timescales when no prior indication in favour of a specific timescale exists. As they also mention, the decision of the timescale should be based on the scientific question to be answered and the expectation of the researcher on how event rate changes as a function of time since study entry or as a function of time since last failure.

Box-Steffensmeier & De Boef (2006) considered also the conditional frailty model in order to account for the event dependence and the heterogeneity among subjects however their work did not include disjoint risk sets as the use of different timescales is not the main subject of their analysis. However, their results do not fall far from ours as they recommend the conditional frailty model as well when one wants to capture both heterogeneity of subjects and event dependence.

While retaining patients in monitoring and treatment has been shown to be linked with improved health outcomes, data from developed and developing countries highlight the difficulties of patient retention. In our study we found that PWID, people with lower education, immigrants from Asia and Africa and people who have an HCV or HBV coinfection are more prone to present gaps in their treatment. In reverse, older patients, EU citizens, MSM, women and people who receive HAART were related with lower risk of presenting gaps in care or monitoring. Hence, the factors who seem to be of more importance for patient retention our results are in line with the results of other studies that were performed in countries that resemble both characteristics of population and characteristics of the epidemic of our setting. Van Beckhoven et al. (2015), tried to explore factors that affect patient retention to care in Belgium, in a cohort that included also a large number of immigrants from the sub-saharan Africa. The findings of another study for assessing factors that lead to patient attrition that is based on the data of the Swiss HIV Cohort Study (SHCS) (Thierfelder et al., 2012) are also in accordance with the findings of our study.

These results point out the groups of people that retention policies should be addressed to, in order to improve the effectiveness of the HIV programs which in its turn will bring us a step closer to the realization of the goals set by UNAIDS for the termination of the pandemic by the year 2030.

# Abstract

The current thesis has three aims. One is to provide the chance of getting familiar with the management of large databases the second is to determine the effect of the timescale that is used in correlated event data and the third is to link factors to patient attrition from HIV monitoring and treatment. For that reason, two different timescales; gap and calendar timescale and a handful of models are implemented, in order to determine which timescale is more appropriate.

Data may be correlated due to multiple events of the same subject. In the case where the study's subjects are experiencing the same type of event multiple times, we refer to the data as recurrent event data. In the analysis of such data, correlation among events and heterogeneity of subjects may present simultaneously in many situations. Therefore we need to find a way to model the association within the observations of a cluster (subject) as well the variance between different subjects (clusters). Variance corrected and shared frailty models can be used for that reason. We also consider the conditional shared frailty model in order to model both correlation and heterogeneity with the use of event-based baseline hazards and random effect. All models were fitted to a dataset of HIV-positive patients in Greece provided by AMACS.

Our simulations showed that the simple frailty models performed slightly better than all other models that were fitted as the conditional frailty models were quite unstable. However the fittings on the empirical data pointed to the opposite direction, as the conditional frailty model under the calendar timescale seemed to be the most appropriate way to assess subject heterogeneity as well as event dependence.

# Appendix

## The stata code

```
set more off

use "atomiko_anamnistiko_aee.dta"
rename AEEDate ExamDate
drop if ExamDate==.
drop if ExamDate<td(01jan1985)
gen Source=1
save "atomiko_anamnistiko_aee.dta", replace
clear

use "atomiko_anamnistiko_emfragma.dta"
rename EmfragmaDate ExamDate
drop if ExamDate==.
gen Source=2
save "atomiko_anamnistiko_emfragma.dta", replace
clear

use "clinic_visits.dta"
rename  DateOfVisit ExamDate
drop if ExamDate==.
replace ExamDate=td(13apr2013) if ExamDate==td(13apr2103)
replace ExamDate=td(01mar2002) if ExamDate==td(01mar2020)
replace ExamDate=td(23sep2014) if ExamDate==td(23sep2024)
replace ExamDate=td(21jul2008) if ExamDate==td(21jul0208)
replace ExamDate=td(17dec2008) if ExamDate==td(17dec0208)
replace ExamDate=td(16sep2008) if ExamDate==td(16sep0208)
replace ExamDate=td(05sep2012) if ExamDate==td(05sep0212)
replace ExamDate=td(30apr2001) if ExamDate==td(30apr0201)
drop if ExamDate==td(02jan1980)
drop if ExamDate==td(01jul1970)
encode Lipoatrofia, gen (Lip)
drop Lipoatrofia
rename Lip Lipoatrofia
rename Enapothesi E
encode E, gen(Enapothesi)
```

```
drop E
gen Source=3
save "clinic_visits.dta", replace
clear


use "exams_aimatologikes.dta"
drop if ExamDate<td(01jan1985)
gen Source=4
save "exams_aimatologikes.dta", replace
clear


use "exams_anosologikes.dta"
drop if ExamDate==.
replace ExamDate=td(17dec2008) if ExamDate==td(17dec0208)
drop if ExamDate<td(01jan1984)
drop if ExamDate>td(01jan2015)
gen Source=5
save "exams_anosologikes.dta", replace
clear


use "exams_bioximikes_new.dta"
drop if ExamDate==.
drop if ExamDate<td(01jan1985)
gen Source=6
save "exams_bioximikes_new.dta", replace
clear

use "exams_iologikes.dta"
drop if ExamDate==.
replace ExamDate=td(30apr2001) if ExamDate==td(30apr0201)
replace ExamDate=td(16sep2008) if ExamDate==td(16sep0208)
replace ExamDate=td(05sep2012) if ExamDate==td(05sep0212)
drop if ExamDate==td(22may1980)
replace ExamDate=td(23sep2014) if ExamDate==td(23sep2024)
replace ExamDate=td(13apr2013) if ExamDate==td(13apr2103)
gen Source=7
```

```
save "exams_iologikes.dta", replace
clear


use "exams_orologikes.dta"
drop if ExamDate==.
drop if ExamDate<td(01jan1985)
gen Source=8
save "exams_orologikes.dta", replace
clear


use "exams_other.dta"
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=9
save "exams_other.dta", replace
clear


use "exams_ourwn.dta"
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=10
save "exams_ourwn.dta", replace
clear


use "hbv_iologikes.dta"
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=11
save "hbv_iologikes.dta", replace
clear
```

```
use "hbv_istology.dta"
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=12
save "hbv_istology.dta", replace
clear


use "hcv_iologikes.dta"
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=13
save "hcv_iologikes.dta", replace
clear


use "hcv_istology.dta"
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=14
save "hcv_istology.dta", replace
clear


use "hiv_resistance.dta"
rename SampleDate ExamDate
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=15
save "hiv_resistance.dta", replace
clear


use "patient_neoplasma.dta"
rename NeoplasmaDate ExamDate
```

```
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
drop if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=16
save "patient_neoplasma.dta", replace
clear


use "patient_other_clinical_state.dta"
rename ClinicalStatusDate ExamDate
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
drop if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=17
save "patient_other_clinical_state.dta", replace
clear


use "patients_category_b.dta"
rename ClinicSymptomDate ExamDate
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=18
save "patients_category_b.dta", replace
clear


use "patients_category_c.dta"
rename NososSyndromDate ExamDate
drop if ExamDate==.
list PatientCode ExamDate if ExamDate<td(01jan1985)
list PatientCode ExamDate if ExamDate>td(01jan2015)
gen Source=19
save "patients_category_c.dta", replace
clear
```

```
*********constructing the atomiko_anamnistiko_new file******************
use "atomiko_anamnistiko.dta"
drop if HypertensionDate==. & StefaniaiaDate==. & DiabitisDate==. ///
& YperlipidaimiaDate==. & LipoatrofiaDate==. & EnapothesiDate==.
save "atomiko_anamnistiko.dta", replace
clear
use "atomiko_anamnistiko.dta"
drop Stef* Diab* Yper* Lip* Enap*
rename HypertensionDate ExamDate
drop if ExamDate==.
save "atomiko_anamnistiko_hyp.dta"
clear
use "atomiko_anamnistiko.dta"
drop Hyp* Diab* Yper* Lip* Enap*
rename StefaniaiaDate ExamDate
drop if ExamDate==.
save "atomiko_anamnistiko_ste.dta"
clear
use "atomiko_anamnistiko.dta"
drop Hyp* Stef* Yper* Lip* Enap*
rename DiabitisDate ExamDate
drop if ExamDate==.
save "atomiko_anamnistiko_diab.dta"
clear
use "atomiko_anamnistiko.dta"
drop Hyp* Stef* Diab*  Lip* Enap*
rename YperlipidaimiaDate ExamDate
drop if ExamDate==.
save "atomiko_anamnistiko_yper.dta"
clear
use "atomiko_anamnistiko.dta"
drop Hyp* Stef* Diab* Yper* Enap*
rename LipoatrofiaDate ExamDate
drop if ExamDate==.
save "atomiko_anamnistiko_lipo.dta"
clear
use "atomiko_anamnistiko.dta"
drop Hyp* Stef* Diab* Yper* Lipo*
```

```
rename EnapothesiDate ExamDate
drop if ExamDate==.
save "atomiko_anamnistiko_enap.dta"
clear
use "atomiko_anamnistiko_hyp.dta"
append using "atomiko_anamnistiko_ste.dta"  ///
"atomiko_anamnistiko_diab.dta" ///
"atomiko_anamnistiko_yper.dta" ///
"atomiko_anamnistiko_lipo.dta" ///
"atomiko_anamnistiko_enap.dta"
gen Source=20
save "atomiko_anamnistiko_new.dta"


***********constructing the overall file for visits**********

use "atomiko_anamnistiko_new.dta"
append using "atomiko_anamnistiko_aee.dta"
"atomiko_anamnistiko_emfragma.dta" ///
"clinic_visits.dta" "exams_aimatologikes.dta" "exams_anosologikes.dta" ///
"exams_bioximikes_new.dta" "exams_iologikes.dta" "exams_orologikes.dta" ///
"exams_other.dta" "exams_ourwn.dta" "hbv_iologikes.dta" ///
"hbv_istology.dta" "hcv_iologikes.dta" "hcv_istology.dta" ///
"hiv_resistance.dta" "patient_neoplasma.dta" ///
"patient_other_clinical_state.dta" "patients_category_b.dta" ///
"patients_category_c.dta"

label var Source "Origin of Data"
label define labSource ///
1 "atomiko_anamnistiko_aee" 2 "atomiko_anamnistiko_emfragma" ///
3 "clinic_visits" 4 "exams_aimatologikes" ///
5 "exams_anosologikes" 6 "exams_bioximikes_new" ///
7 "exams_iologikes" 8 "exams_orologikes" ///
9 "exams_other" 10 "exams_ourwn" ///
11 "hbv_iologikes" 12 "hbv_istology" ///
13 "hcv_iologikes" 14 "hcv_istology" ///
15 "hiv_resistance" 16 "patient_neoplasma" ///
17 "patient_other_clinical_state" 18 "patients_category_b" ///
19 "patients_category_c" 20 "atomiko anamnistiko"
```

```
label values Source labSource

label var Enapothesi "Fat deposition"
label var Lipoatrofia "Lipoatrophy"
label var ExamDate "Date of visit"
label var Yperlipidaimia "Hyperlipidemia"
label var Diabitis "Diabetes"
label var Stefaniaia "Coronary artery disease"

drop if ExamDate<td(01jan1985)
save "visits_overall.dta",clear

************constructing the dateofdeath file**************
use "hospitalization.dta"
keep if Ekbasi==1
*95 remaining observations
drop EntryDate Diagnosis Ekbasi
rename ExitDate DateofDeath
label var DateofDeath "Date of Death"
gen State=1
save "hospitalization_death.dta"
clear
use "last_state.dta"
drop LastState Lost2FollowUp LastKnownT NewClinic WithdrawalDate Immediate
Contributing1 Contributing2 Contributing3 Contributing4 Underlying Notes
gen State=1
drop if DeathDate==. & DateDeathKnown==.
*(6595 observations deleted) leaving 635 observations
gen D=DeathDate
format D %td
replace D=DateDeathKnown if D==.
label var D "Date of Death"
drop DeathDate DateDeathKnown
save "last_state_death.dta"
bysort PatientCode: gen sort=_n
keep if sort==1
*(43 observations deleted)
merge 1:1 PatientCode using "hospitalization_death.dta"
replace DateofDeath=D if DateofDeath==.
```

```
li if DateofDeath<td(01jan1985)
drop if if DateofDeath<td(01jan1985)
*(1 observation deleted) leaving 593 observations
drop D sort _merge
label define lab 1 "dead"
label values State lab
rename fromCenter fromCenter_dod
save "dateofdeath.dta"
clear

*********merging dateofdeath file with demographic_data***************
use "demographic_data.dta"
merge 1:1 PatientCode using "dateofdeath.dta"
drop _merge
save "demographic_data.dta", replace
clear

*********merging visit_overall file with demographic_data**************
use "visits_overall.dta"
merge m:1 PatientCode using "demographic_data.dta"
**matched 1,983,108
drop _merge
label var PatientCode "Unique Patient Code"
label var ExamDate "Date of biochemical examination"
label var Code "Code of biochemical test"
label var Value "Result of biochemical test"
label var Lower "Lower limit of biochemical test"
label var Upper "Upper limit of biochemical test"
label var Unit "Measuring unit of biochemical test"
label var BirthDate "Date of Birth"
label var fromCenter "Code of Clinic"
save "complete_data.dta"
clear


************************************************************************
use "complete_data.dta"
sort PatientCode ExamDate
gen sortgen=_n
```

```
label var sortgen "sorted by Patient and Date of visit"

*generating ascending serial number of visit for every patient
bysort PatientCode ExamDate: gen sort=_n

by PatientCode: egen mode=mode(fromCenter)
label var mode "Most visited Clinic per patient"

*****removing multiple lines for the same visit
keep if sort==1
drop sort

sort PatientCode ExamDate
replace sortgen=_n

*generating variables for the beginning and end of follow-up period
gen tstart=.
format tstart %td
gen tstop=.
format tstop %td

label var tstart "Begin of risk period"
label var tstop "End of risk period"

*generating new ascending serial number of visit for every patient
sort PatientCode ExamDate
by PatientCode: gen sort3=_n
label var sort3 "ascending serial number of visit per patient"

*generating descending serial number of visit for every patient
gen sort4=-sort3
label var sort4 "negative sorting of visits per patient"
sort PatientCode sort4
by PatientCode: gen desc_sort=_n
label var desc_sort "descending serial number of visit per patient"

by PatientCode: gen totvisits=_N
label var totvisits "Total number of visits"
```

```
replace tstart=ExamDate
sort sortgen
replace tstop=ExamDate
bysort PatientCode: replace tstop=tstop[_n+1]

drop MELCode fromCenter
rename BirthDate dateofbirth
merge m:1 PatientCode using "patients.dta"

*taking the last known examination date by Center as the database closure
bysort fromCenter: egen Studyclosure=max(ExamDate)
gen d2=Studyclosure-ExamDate if desc_s==1

save "data4split.dta", replace
clear

************************************
***for those with one visit (#66)***
************************************
use  "data4split.dta"
keep if totvisits==1
*(231971 observations deleted) 66 remaining
replace tstop=DateofDeath if DateofDeath-ExamDate<365.25
*3 died within a year after their 1st visit
replace tstop=Studyclosure if Studyclosure-ExamDate<365.25
*18 had a visit within a year before study close
gen gapyear=1 if Studyclosure-ExamDate>=365.25 & DateofDeath==.
replace tstop=ExamDate+182.625 if Studyclosure-ExamDate>=365.25 & DateofDeath==.
*(45 real changes made)
replace gapyear=0 if gapyear==.
save "thosewith1visit.dta", replace
clear

************************************
**** #13 with postmortem visits ****
************************************

use "data4split.dta"
drop if totvisits==1
```

```
*(66 observations deleted)
gen dif=DateofDeath-ExamDate if desc_s==1
label var dif "Difference between date of death and visit date"
keep if Pat==201319013 | Pat==3413| Pat==6061| Pat==4525|Pat==1242 ///
|Pat==1065|Pat==7918|Pat==1943|Pat==202012907|Pat==200783461 ///
|Pat==201678154|Pat==5067|Pat==202902092
save "ThoseWithPostmortemVisits13.dta", replace
clear

************************************
****** creating gap  variable ******
******  & filling last tstop  ******
************************************

use "data4split.dta"
sort PatientCode ExamDate
drop if totvisits==1
gen dif=DateofDeath-ExamDate if desc_s==1
label var dif "Difference between date of death and visit date"
drop if Pat==201319013 | Pat==3413| Pat==6061| Pat==4525|Pat==1242 ///
|Pat==1065|Pat==7918|Pat==1943|Pat==202012907|Pat==200783461 ///
|Pat==201678154|Pat==5067|Pat==202902092

replace tstop=DateofDeath if dif>0 & dif<365.25
gen gapyear=1 if dif>365.25 & dif!=.
replace tstop=ExamDate+182.625 if dif>365.25 & dif!=.
replace tstop=DateofDeath+1 if dif==0

*for those who have at least 2 visits totally and a visit within the year of
2014
replace tstop=Studyclosure if desc==1 & d2<365.25 & DateofDeath==.

*for those who have at least 2 visits totally but without a visit within the
year of 2014
replace gapyear=1 if desc==1 & d2>=365.25 & DateofDeath==.
replace tstop=ExamDate+182.625 if desc==1 & d2>=365.25 & DateofDeath==.

replace gapyear=1 if tstop-ExamDate>365.25
replace gapyear=0 if gapyear==.
```

```
label define gapyearlab 0 "no" 1 "yes"
label values gapyear gapyearlab
drop if ExamDate<td(01jan1996)

*serial number of gaps by each patient
by PatientCode : generate num96=sum( gapyear ) if gapyear ==1
label var num96 "ascending serial number of gaps"

*total number of gaps per patient 1996-2014
egen totalnum=max(num96), by(PatientCode)
label var totalnum "total number of gaps 1996-2014"
replace totalnum=0 if totalnum==.

***First ever visit for every Patient after 96
sort PatientCode ExamDate
by PatientCode: gen sort6=_n
label var sort6 "ascending serial number of visit per patient after 96"

bysort PatientCode (sort6): gen firstvisit96=ExamDate if sort6==1
bysort PatientCode (sort6): carryforward firstvisit96,replace
format firstvisit96 %td

***First visit after every gap
forvalues i=1/8 {
bysort PatientCode (sort6): gen fvag`i'=ExamDate[_n+1] if num==`i' & desc_sort!
=1
bysort PatientCode (sort6): carryforward fvag`i',replace
format fvag`i' %td
}

label var fvag1 "First Visit After 1 Gap"
label var fvag2 "First Visit After 2 Gap"
label var fvag3 "First Visit After 3 Gap"
label var fvag4 "First Visit After 4 Gap"
label var fvag5 "First Visit After 5 Gap"
label var fvag6 "First Visit After 6 Gap"
label var fvag7 "First Visit After 7 Gap"
label var fvag8 "First Visit After 8 Gap"
```

```
by PatientCode: gen totvisits1996_2014=_N
label var totvisits "Total number of visits 1996-2014"

save "1996-2014gaps.dta",replace
clear

************************************
****** those who have NO gaps ******
************* # 102306 **************
************************************
use "1996-2014gaps.dta"
keep if totalnum==0
sort sortgen
replace tstart=firstvisit96 if desc_sort==1
keep if desc_sort==1
save "ThoseWhoHave0gaps.dta",replace
clear

************************************
******* those who have 1 gap *******
************* # 79699 **************
************************************
use "1996-2014gaps.dta"
keep if totalnum==1
count if gapyear==desc_sort==1
codebook Pat if State!=.

***for those who had their gap at their last visit #1601
replace tstart=firstvisit96 if desc_sort==1 & num96==1

***for those who had NOT their gap at their last visit #1167
replace tstart=firstvisit96 if desc_sort!=1 & num96==1
replace tstop=ExamDate+182.625 if desc_sort!=1 & num96==1
replace tstart=fvag1 if num96==. & desc_sort==1
replace tstop=Studyclosure if num96==. & desc_sort==1 & Studyclosure-
ExamDate<365.25
replace tstop=DateofDeath if num96==. & desc_sort==1 & tstop==DateofDeath+1
replace tstop=DateofDeath if num96==. & desc_sort==1 & DateofDeath-
ExamDate<365.25 & DateofDeath-Studyclosure<0
```

```
keep if (num==1 & desc_sort!=1) | (num==. & desc_sort==1) | (num==1 & desc_sort==1)

save "ThoseWhoHave1gap.dta",replace
clear

************************************
******* those who have 2 gap *******
************* # 26392 **************
************************************
use "1996-2014gaps.dta"
keep if totalnum==2

*for the 1st gap #898
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

***for those who had their 2nd gap at their last visit #524
replace tstart=fvag1 if num96==2 & desc_sort==1

***for those who had their 2nd gap BEFORE their last visit #374
replace tstart=fvag1 if num96==2 & desc_sort!=1
replace tstop=ExamDate+182.625 if num==2 & desc_sort!=1

replace tstart=fvag2 if num96==. & desc_sort==1
replace tstop=Studyclosure if num96==. & desc_sort==1 & Studyclosure-ExamDate<365.25
replace tstop=DateofDeath if num96==. & desc_sort==1 & tstop==DateofDeath+1
replace tstop=DateofDeath if num96==. & desc_sort==1 & DateofDeath-ExamDate<365.25 & DateofDeath-Studyclosure<0

keep if (num96==1) | (num96==2 & desc_sort!=1) | (num96==. & desc_sort==1) | (num96==2 & desc_sort==1)

save "ThoseWhoHave2gaps.dta",replace
clear

************************************
```

```
******* those who have 3 gap *******
************* # 9578  **************
************************************
use "1996-2014gaps.dta"
keep if totalnum==3

*for the 1st gap #324
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

*for the 2d gap #324
replace tstart=fvag1 if num96==2
replace tstop=ExamDate+182.625 if num96==2

***for those who had their 3rd gap at their last visit #194
replace tstart=fvag2 if num96==3 & desc_sort==1
count if num96==3 & desc_sort==1 & tstop==ExamDate+182.625

***for those who had their 3rd gap BEFORE their last visit #130
replace tstart=fvag2 if num96==3 & desc_sort!=1
replace tstop=ExamDate+182.625 if num==3 & desc_sort!=1

replace tstart=fvag3 if num96==. & desc_sort==1
replace   tstop=Studyclosure   if   num96==.   &   desc_sort==1   &   Studyclosure-
ExamDate<365.25
replace tstop=DateofDeath if num96==. & desc_sort==1 & tstop==DateofDeath+1
replace   tstop=DateofDeath   if   num96==.   &   desc_sort==1   &   DateofDeath-
ExamDate<365.25 & DateofDeath-Studyclosure<0

keep  if  (num96==1)  |(num96==2)  |  (num96==3  &  desc_sort!=1)  |  (num96==.  &
desc_sort==1) | (num96==3 & desc_sort==1)

save "ThoseWhoHave3gaps.dta",replace
clear

************************************
******* those who have 4 gap *******
************* # 4643  **************
************************************
```

```
use "1996-2014gaps.dta"
keep if totalnum==4

*for the 1st gap #148
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

*for the 2nd gap #148
replace tstart=fvag1 if num96==2
replace tstop=ExamDate+182.625 if num96==2

*for the 3rd gap #148
replace tstart=fvag2 if num96==3
replace tstop=ExamDate+182.625 if num96==3

***for those who had their 4th gap at their last visit #90
replace tstart=fvag3 if num96==4 & desc_sort==1

***for those who had their 4th gap BEFORE their last visit #58
replace tstart=fvag3 if num96==4 & desc_sort!=1
replace tstop=ExamDate+182.625 if num==4 & desc_sort!=1

replace tstart=fvag4 if num96==. & desc_sort==1
replace tstop=Studyclosure if num96==. & desc_sort==1 & Studyclosure-
ExamDate<365.25
replace tstop=DateofDeath if num96==. & desc_sort==1 & DateofDeath-
ExamDate<365.25 & DateofDeath-Studyclosure<0

keep if (num96==1) |(num96==2) | (num96==3) |(num96==4 & desc_sort!=1) |
(num96==. & desc_sort==1) | (num96==4 & desc_sort==1)

save "ThoseWhoHave4gaps.dta",replace
clear

************************************
******* those who have 5 gap *******
************* # 1419  **************
************************************
use "1996-2014gaps.dta"
```

```
keep if totalnum==5

*for the 1st gap #
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

*for the 2nd gap #
replace tstart=fvag1 if num96==2
replace tstop=ExamDate+182.625 if num96==2

*for the 3rd gap #
replace tstart=fvag2 if num96==3
replace tstop=ExamDate+182.625 if num96==3

*for the 4th gap #
replace tstart=fvag3 if num96==4
replace tstop=ExamDate+182.625 if num96==4

***for those who had their 5th gap at their last visit #23
replace tstart=fvag4 if num96==5 & desc_sort==1

***for those who had their 5th gap BEFORE their last visit #23
replace tstart=fvag4 if num96==5 & desc_sort!=1
replace tstop=ExamDate+182.625 if num96==5 & desc_sort!=1

replace tstart=fvag5 if num96==. & desc_sort==1
replace tstop=Studyclosure if num96==. & desc_sort==1 & Studyclosure-ExamDate<365.25
replace tstop=DateofDeath if num96==. & desc_sort==1 & DateofDeath-ExamDate<365.25 & DateofDeath-Studyclosure<0

keep if (num96==1) |(num96==2) | (num96==3) | (num96==4) |(num96==5 & desc_sort!=1) | (num96==. & desc_sort==1) | (num96==5 & desc_sort==1)

save "ThoseWhoHave5gaps.dta",replace
clear

************************************
******* those who have 6 gap *******
```

```
************  # 611  *************
**********************************
use "1996-2014gaps.dta"
keep if totalnum==6

*for the 1st gap #
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

forvalues i=1/5 {
replace tstart=fvag`i' if num96==`i'+1
replace tstop=ExamDate+182.625 if num96==`i'+1
}

***for those who had their 6th gap at their last visit #16
replace tstart=fvag5 if num96==6 & desc_sort==1

***for those who had their 6th gap BEFORE their last visit #8
replace tstart=fvag5 if num96==6 & desc_sort!=1
replace tstop=ExamDate+182.625 if num96==6 & desc_sort!=1

replace tstart=fvag6 if num96==. & desc_sort==1
replace tstop=Studyclosure if num96==. & desc_sort==1 & Studyclosure-
ExamDate<365.25
replace tstop=DateofDeath if num96==. & desc_sort==1 & DateofDeath-
ExamDate<365.25 & DateofDeath-Studyclosure<0

keep if (num96==1) | (num96==2) | (num96==3) | (num96==4) | (num96==5) |
(num96==6 & desc_sort!=1) | (num96==. & desc_sort==1) | (num96==6 &
desc_sort==1)

save "ThoseWhoHave6gaps.dta",replace
clear

**********************************
******* those who have 7 gap *******
************  # 214  *************
**********************************
use "1996-2014gaps.dta"
```

```
keep if totalnum==7

*for the 1st gap #
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

forvalues i=1/6 {
replace tstart=fvag`i' if num96==`i'+1
replace tstop=ExamDate+182.625 if num96==`i'+1
}

***for those who had their 7th gap at their last visit #4
replace tstart=fvag6 if num96==7 & desc_sort==1
count if num96==7 & desc_sort==1 & tstop==ExamDate+182.625

***for those who had their 7th gap BEFORE their last visit #8
replace tstart=fvag7 if num96==. & desc_sort==1
replace tstop=Studyclosure if num96==. & desc_sort==1 & Studyclosure-ExamDate<365.25

keep if (num96==1) | (num96==2) | (num96==3) | (num96==4) | (num96==5) | (num96==6) | (num96==7 & desc_sort!=1) | (num96==. & desc_sort==1) | (num96==7 & desc_sort==1)

save "ThoseWhoHave7gaps.dta",replace
clear

************************************
******* those who have 8 gap *******
*************  # 19  ***************
************************************
use "1996-2014gaps.dta"
keep if totalnum==8

*for the 1st gap #
replace tstart=firstvisit96 if num96==1
replace tstop=ExamDate+182.625 if num96==1

forvalues i=1/7 {
```

```
replace tstart=fvag`i' if num96==`i'+1
replace tstop=ExamDate+182.625 if num96==`i'+1
}

keep if (num96==1) | (num96==2) | (num96==3) | (num96==4) | (num96==5) |
(num96==6) | (num96==7) | (num96==8)

save "ThoseWhoHave8gaps.dta",replace
clear

***********************************************************
*****************appending the datasets********************
***********************************************************

use "ThoseWhoHave0gaps.dta"
append using "ThoseWhoHave1gap.dta"  ///
"ThoseWhoHave2gaps.dta" ///
"ThoseWhoHave3gaps.dta" ///
"ThoseWhoHave4gaps.dta" ///
"ThoseWhoHave5gaps.dta" ///
"ThoseWhoHave6gaps.dta" ///
"ThoseWhoHave7gaps.dta" ///
"ThoseWhoHave8gaps.dta"

drop sort6 dif d2 Studyclosure _merge totvisits desc_sort sort4 sort3
fromCenter_dod OtherClinic OtherDrugName OtherDrugValue PressureSystolic
PressureDiastolic AbsoluteCD4 PercentCD4 AbsoluteCD8 PercentCD8 Name Surname
CodeType Result Operator Units Method Type ExtraField ExtraValue

save "All_gaps.dta",replace
clear

************************************************************
use "All_gaps.dta"

keep fvag1 fvag2 fvag3 fvag5 fvag4 fvag6 fvag7 fvag8 PatientCode ExamDate Source
EnrollDate Race Sex Education State DateofDeath mode tstart tstop fromCenter
gapyear num96 totalnum totvisits1996_2014 firstvisit96
```

```
gen FUtime=tstop-tstart
label var FUtime "Follow-up time"

replace tstop=tstop+1 if PatientCode==999540 & FUtime==0
replace FUtime=1 if FUtime==0

bysort PatientCode (ExamDate): gen sort7=_n
replace State=2 if State==.
label define Statelbl 1 "dead" 2 "alive"
label values State Statelbl
label var State "Is the patient dead or alive"
label var fromCenter "Reference Clinic"
label var gapyear "Failure indicator"
label var totvisits1996_2014 "total number of visits 1996-2014"
label var sort "ascending number of risk periods per patient"

save "All_gaps.dta",replace
clear

*********************firstexamdate + tstart********************************

use "All_gaps.dta"
rename firstvisit96 firstexamdate
keep if sort7==1
keep Pat first
save "firstexamdate.dta"
clear

use "All_gaps.dta"
keep PatientCode tstart sort7
save "tstart.dta"
clear

*********************DateofBirth & age******************
use "patients.dta"
keep PatientCode BirthDate
rename BirthDate dob
label var dob "Date of Birth"
merge 1:1 PatientCode using "firstexamdate.dta"
```

```
drop if _merge==1
drop _merge
gen age= firstexamdate-dob
drop if age==.
label var age "Age at first appearance"
replace age=age/365.25
save "dob_age.dta",replace
clear

*******************origin************************
use "demographic_data.dta"
gen origin=4 if Origin==0 | Origin==99 | Origin==255
replace origin=1 if Origin==10 | Origin==11 | Origin==12
replace origin=2 if Origin==20 | Origin==40 | Origin==60
replace origin=3 if Origin==50 | Origin==52
replace origin=0 if Origin==70 | Origin==71 | Origin==72
label drop originlbl
label define originlbl 4 "Not specified" 1"Africa" 2 "Asia & Australia" 3
"Americas" 0 "Europe"
label values origin originlbl
save  "origin.dta",replace
clear

*******************coinfections******************

use "exams_orologikes.dta"
merge m:1 PatientCode using "firstexamdate.dta"
drop if ExamDate<td(01jan1996)
drop if _merge==1 |_merge==2
encode Type,gen(type)
keep if type==7 | type==11
sort Pat ExamDate
encode Result, gen(result)
recode result (2=0)
label drop result
label define resultlbl 0 "-" 1 "+"
label values result resultlbl

****HCV status for those who were tested positive within a year from their first
```

```
exam date
by PatientCode (ExamDate): egen sumHCV=sum(result) if type==7 &
ExamDate<=firstexamdate+365.25
by PatientCode (ExamDate): gen sumHCV2=sum(result) if type==7 &
ExamDate<=firstexamdate+365.25


****HCV status for those who were tested positive after a year from their first
exam date but never tested negative throughout their follow up time
by PatientCode (ExamDate): gen hcvindicator=-100 if result==0 & type==7 &
ExamDate>firstexamdate+365.25
by PatientCode (ExamDate): replace hcvindicator=1 if result==1 & type==7 &
ExamDate>firstexamdate+365.25
by PatientCode (ExamDate): gen sumHCV3=sum(hcvindicator)

gen HCV=1 if sumHCV2!=0 & sumHCV2!=.
replace HCV=1 if sumHCV3>0 & sumHCV3!=.

****HBV status for those who were tested positive within a year from their first
exam date
by PatientCode (ExamDate): gen sumHBV=sum(result) if type==11 &
ExamDate<=firstexamdate+365.25


****HBV status for those who were tested positive after a year from their first
exam date but never tested negative throughout their follow up time
by PatientCode (ExamDate): gen hbvindicator=-100 if result==0 & type==11 &
ExamDate>firstexamdate+365.25
by PatientCode (ExamDate): replace hbvindicator=1 if result==1 & type==11 &
ExamDate>firstexamdate+365.25
by PatientCode (ExamDate): gen sumHBV2=sum(hbvindicator)

gen HBV=1 if sumHBV!=0 & sumHBV!=.
replace HBV=1 if sumHBV2>0 & sumHBV2!=.

bysort PatientCode (ExamDate): gen sort=_n
gen negsort=-sort
bysort PatientCode (negsort): carryforward HCV,gen(HCV2)
bysort PatientCode (negsort): carryforward HBV,gen(HBV2)
```

```
keep if sort==1
keep if HCV2!=. | HBV2!=.
keep PatientCode HCV2 HBV2
rename HCV2 HCV
rename HBV2 HBV
replace HCV=0 if HCV==.
replace HBV=0 if HBV==.

save "coinfections.dta",replace
clear

********************virus load longitudinal********************
use "exams_iologikes.dta"

sort PatientCode ExamDate
drop if ExamDate<td(01jan1996)
by PatientCode (ExamDate):gen sort=_n

encode Operator,gen(Oper)
recode Oper (1=-1)(2=0)(3=1)
label define Operlbl -1 "<" 0 "=" 1">"
label values Oper Operlbl
drop Operator
rename Oper hivRNAoperator

forvalues i=1/5 {
replace Value=Value[_n+`i'] if Value==0 & PatientCode==PatientCode[_n+`i'] &
ExamDate[_n+`i']<ExamDate+365.25 & Value[_n+`i']!=0
replace hivRNAoperator=hivRNAoperator[_n+`i'] if Value==0 &
PatientCode==PatientCode[_n+`i'] & ExamDate[_n+`i']<ExamDate+365.25 &
Value[_n+`i']!=0
}

replace hivRNAoperator=hivRNAoperator[_n-1] if Value==0 &
PatientCode==PatientCode[_n-1] & ExamDate<ExamDate[_n-1]+365.25 & Value[_n-1]!=0
replace Value=Value[_n-1] if Value==0 & PatientCode==PatientCode[_n-1] &
ExamDate<ExamDate[_n-1]+365.25 & Value[_n-1]!=0
```

```
replace hivRNAoperator=0 if (PatientCode==5492 & sort==1) |
(PatientCode==203123632 & sort==1)

append using "tstart.dta"
replace ExamDate=tstart if tstart!=.
gen ts2=tstart
format ts2 %td
gen value=Value

**replace missing values with non missing within a window of 1 year
forvalues x=1/8 {
forvalues i=1/10 {
bysort Pat (Exam sort7):replace hivRNAoperator=hivRNAoperator[_n+`i'] if
missing(value) & sort7==`x' & value[_n+`i']!=. & ExamDate<ts2+365.25 &
PatientCode==PatientCode[_n+`i']
bysort Pat (Exam sort7):replace value=value[_n+`i'] if missing(value) &
sort7==`x' & value[_n+`i']!=. & ExamDate<ts2+365.25 &
PatientCode==PatientCode[_n+`i']
}
}

forvalues x=1/8 {
replace hivRNAoperator=hivRNAoperator[_n-1] if missing(value) & sort7==`x' &
value[_n-1]!=. & ts2<ExamDate+365.25 & PatientCode==PatientCode[_n-1]
replace value=value[_n-1] if missing(value) & sort7==`x' & value[_n-1]!=. &
ts2<ExamDate+365.25 & PatientCode==PatientCode[_n-1]
}

drop if Pat==176 | Pat==836 | Pat==1047 | Pat==1425 | Pat==1590 | Pat==1812
keep if sort7!=. & value!=.
keep PatientCode value tstart hivRNAoperator sort7

replace value=value/2 if hivRNAoperator==-1
rename value virusload
label var virusload "Virus load"
keep PatientCode tstart sort7 virusload
gen logvirusload=log10(virusload)
label var logvirusload "base-10 logarithm of Virus Load"
```

```
save "virusload_longitudinal.dta",replace
clear

***********************CD4 count longitudinal*******************************
use "exams_anosologikes.dta"

keep PatientCode ExamDate AbsoluteCD4

append using "tstart.dta"
sort Pat Exa
drop if Exam<td(01jan1996)
bysort Pat (Exam):gen sort=_n
replace ExamDate=tstart if tstart!=.

sort Pat Exam sort7
gen ts2=tstart
format ts2 %td

gen cd4=AbsoluteCD4
**change all the zero values to missing values
replace cd4=. if cd4==0

**replace missing values with non missing within a window of 1 year
forvalues x=1/8 {
forvalues i=1/10 {
bysort Pat (Exam sort7):replace cd4=cd4[_n+`i'] if missing(cd4) & sort7==`x' & cd4[_n+`i']!=. & ExamDate<ts2+365.25 & PatientCode==PatientCode[_n+`i']
}
}

forvalues x=1/8 {
replace cd4=cd4[_n-1] if missing(cd4) & sort7==`x' & cd4[_n-1]!=. & ts2<ExamDate+365.25 & PatientCode==PatientCode[_n-1]
}

drop if Pat==999568 | Pat==200313005 | Pat==200904869 | Pat==200914354 | Pat==201054938 | Pat==201060641 | Pat==201153724 | Pat==201250628 | Pat==201315217| Pat==201395036| Pat==201446346| Pat==201486253| Pat==201605947 | Pat==201896852 | Pat==201986275 | Pat==202100380 | Pat==202115600 |
Pat==202226169 | Pat==202235721 | Pat==202308329 | Pat==202310241
keep if cd4!=. & sort7!=.

gen CD4log=log(cd4)
label var CD4log "longitudinal natural logarithm of CD4 count"
keep PatientCode tstart sort7 CD4log cd4
gen cd4cat=0 if cd4>=350
replace cd4cat=1 if cd4<350 & cd4>=200
replace cd4cat=2 if cd4<200
label define cd4catlab 0 "CD4>=350" 1 "CD4>=200 & <350" 2 "CD4<200"
save "CD4_longitudinal.dta",replace
clear

****************Possible source of infection********************

use "demographic_data.dta"
gen psoi=4 if PossibleSourceInfection==0 | PossibleSourceInfection==8 |
PossibleSourceInfection==9
replace psoi=1 if PossibleSourceInfection==1 | PossibleSourceInfection==3
replace psoi=2 if PossibleSourceInfection==2
replace psoi=3 if PossibleSourceInfection==4 | PossibleSourceInfection==5 |
PossibleSourceInfection==6
replace psoi=0 if PossibleSourceInfection==7

label define psoilbl 4 "Not-specified" 1 "Homosexual/Bisexual contact" 2 "IDU" 3
"Blood transfusion" 0 "Heterosexual contact"
label values psoi psoilbl
label var psoi "Possible source of infection"
keep PatientCode psoi PossibleSourceInfection
save "possource.dta",replace
clear

*******************education***********************
use "demographic_data.dta"

encode Education, gen(edu)
gen education=0 if edu==1 | edu==5 | edu==10
replace education=1 if edu==6 | edu==7
replace education=2 if edu==3 | edu==4
replace education=3 if edu==2 | edu==8
replace education=4 if edu==9 |edu==.
label define educationlbl 0 "None / Primary Education" 1 "Secondary Education" 2
"Higher / Vocational Education" 3 "Academical Education" 4 "Not Specified"
label values education educationlbl
keep PatientCode edu education
label var education "Level of Education"

save "education.dta",replace
clear

*********************HAART**********************
use "haart.dta"

gen Haart=0 if HAART==0
replace Haart=1 if HAART!=0
sort PatientCode StartDate
drop if StartDate<td(01jan1996)
bysort PatientCode (StartDate): gen sort=_n
gen Haartindicator=-1 if Haart==0
replace Haartindicator=100 if Haart==1
bysort PatientCode (StartDate): egen sumindicator2=sum(Haartindicator)
keep if sort==1
gen Haart=0 if sumindicator2<0
replace Haart=1 if sumindicator2>0

label define Haartlbl 0 "NO" 1 "YES"
label values Haart Haartlbl
label var Haart "Ever received HAART"
drop sumindicator Haartindicator Haart fromCenter Compliance Schema

save "Haart(ever).dta",replace
clear

*****************************************************
use "All_gaps.dta"

merge m:1 PatientCode using "Haart(ever).dta"
replace Haart=0 if Haart==.
```

```
drop _merge

merge m:1 PatientCode using "education.dta"
drop if _merge==2
drop _merge

merge m:1 PatientCode using "possource.dta"
drop if _merge==2
drop _merge

merge m:1 PatientCode using "dob_age.dta"
drop if _merge==2
drop _merge

merge m:1 PatientCode using "origin.dta"
drop if _merge==2
drop _merge

merge m:1 PatientCode using "coinfections.dta"
replace HCV=0 if HCV==.
replace HBV=0 if HBV==.
gen hcvhbv=1 if HCV==1 | HBV==1
replace hcvhbv=0 if hcvhbv==.
drop _merge

merge 1:1 PatientCode tstart using "CD4_longitudinal.dta"
drop _merge

merge 1:1 PatientCode tstart using "virusload_longitudinal.dta"
drop _merge

gen sex=Sex if Sex!=3
recode sex (1=0)(2=1)
label drop sexlbl
label define sexlbl 0 "male" 1 "female"
label values sex sexlbl
gen gap=gapyear

save "Complete.dta",replace
```

```
drop if sex==.
drop if age<18
keep PatientCode gap sex hcvhbv logvirusload CD4log cd4 cd4cat origin FUtime
Haart tstart tstop education psoi age
drop if CD4log==.
drop if logvirusload==.
drop if age==.

saveold "!!gia tin R.dta",replace
```

## The R code

```
my.summary.coxph.penal <-
  function (object, conf.int = 0.95, scale = 1, terms = FALSE,
            maxlabel = 25, digits = max(options()$digits - 4, 3), ...)
  {
    if (!is.null(object$call)) {
      cat("Call:\n")
      dput(object$call)
      cat("\n")
    }
    if (!is.null(object$fail)) {
      cat(" Coxreg failed.", object$fail, "\n")
      return()
    }
    savedig <- options(digits = digits)
    on.exit(options(savedig))
    omit <- object$na.action
    if (length(omit))
      cat("  n=", object$n, " (", naprint(omit), ")\n", sep = "")
    else cat("  n=", object$n, "\n")
    coef <- object$coef
    if (length(coef) == 0 && length(object$frail) == 0)
      stop("Penalized summary function can't be used for a null model")
    if (length(coef) > 0) {
      nacoef <- !(is.na(coef))
      coef2 <- coef[nacoef]
      if (is.null(coef) | is.null(object$var))
        stop("Input is not valid")
      se <- sqrt(diag(object$var))
    }
    pterms <- object$pterms
    nterms <- length(pterms)
    npenal <- sum(pterms > 0)
    print.map <- rep(0, nterms)
    if (!is.null(object$printfun)) {
      temp <- unlist(lapply(object$printfun, is.null))
      print.map[pterms > 0] <- (1:npenal) * (!temp)
    }
    print1 <- NULL
    pname1 <- NULL
    if (is.null(object$assign2))
      alist <- object$assign[-1]
    else alist <- object$assign2
    print2 <- NULL
    for (i in 1:nterms) {
      kk <- alist[[i]]
      if (print.map[i] > 0) {
        j <- print.map[i]
        if (pterms[i] == 2)
          temp <- (object$printfun[[j]])(object$frail,
                                          object$fvar, , object$df[i],
object$history[[j]])
        else temp <- (object$printfun[[j]])(coef[kk], object$var[kk,
                                                                 kk],
object$var2[kk, kk], object$df[i], object$history[[j]])
        print1 <- rbind(print1, temp$coef)
        if (is.matrix(temp$coef)) {
          xx <- dimnames(temp$coef)[[1]]
          if (is.null(xx))
            xx <- rep(names(pterms)[i], nrow(temp$coef))
          else xx <- paste(names(pterms)[i], xx, sep = ", ")
          pname1 <- c(pname1, xx)
        }
        else pname1 <- c(pname1, names(pterms)[i])
        print2 <- c(print2, temp$history)
      }
      else if (terms && length(kk) > 1) {
        pname1 <- c(pname1, names(pterms)[i])
        temp <- coxph.wtest(object$var[kk, kk], coef[kk])$test
        print1 <- rbind(print1, c(NA, NA, NA, temp, object$df[i],
                                  1 - pchisq(temp, 1)))
      }
      else {
        pname1 <- c(pname1, names(coef)[kk])
        tempe <- (diag(object$var))[kk]
        temp <- coef[kk]^2/tempe
        print1 <- rbind(print1, cbind(coef[kk], sqrt(tempe),
                                      sqrt((diag(object$var2))[kk]), temp, 1, 1
- pchisq(temp,

1)))
      }
    }
    temp <- cbind(format(print1[, 1]), format(print1[, 2]), format(print1[,
                                                                          3]),
format(round(print1[, 4], 2)), format(round(print1[,

5], 2)), format(signif(print1[, 6], 2)))
    temp <- ifelse(is.na(print1), "", temp)
    dimnames(temp) <- list(substring(pname1, 1, maxlabel), c("coef",
                                                             "se(coef)", "se2",
"Chisq", "DF", "p"))
    prmatrix(temp, quote = FALSE)
    if (conf.int & length(coef) > 0) {
      z <- qnorm((1 + conf.int)/2, 0, 1)
      coef <- coef * scale
      se <- se * scale
      tmp <- cbind(exp(coef), exp(-coef), exp(coef - z * se),
                   exp(coef + z * se))
      dimnames(tmp) <- list(substring(names(coef), 1, maxlabel),
                            c("exp(coef)", "exp(-coef)", paste("lower .",
round(100 *

conf.int, 2), sep = ""), paste("upper .", round(100 *

conf.int, 2), sep = "")))
```

```
      cat("\n")
      prmatrix(tmp)
    }
    logtest <- -2 * (object$loglik[1] - object$loglik[2])
    sctest <- object$score
    cat("\nIterations:", object$iter[1], "outer,", object$iter[2],
        "Newton-Raphson\n")
    if (length(print2)) {
      for (i in 1:length(print2)) cat("   ", print2[i], "\n")
    }
    if (is.null(object$df))
      df <- sum(!is.na(coef))
    else df <- round(sum(object$df), 2)
    cat("Degrees of freedom for terms=", format(round(object$df,
                                                       1)), "\n")
    cat("Rsquare=", format(round(1 - exp(-logtest/object$n),
                                  3)), "  (max possible=", format(round(1 - exp(2
* object$loglik[1]/object$n),
                                                                        3)),
")\n")
    cat("Likelihood ratio test= ", format(round(logtest, 2)),
        "  on ", df, " df,", "   p=", format(1 - pchisq(logtest,
                                                         df)), "\n", sep = "")
    if (!is.null(object$wald.test))
      cat("Wald test            = ", format(round(object$wald.test,
                                                   2)), "  on ", df, " df,   p=",
format(1 - pchisq(object$wald.test,

df)), sep = "")
    if (!is.null(object$score))
      cat("\nScore (logrank) test = ", format(round(sctest,
                                                     2)), "  on ", df, " df,", "
p=", format(1 - pchisq(sctest,

df)), sep = "")
    if (is.null(object$rscore))
      cat("\n")
    else cat(",   Robust = ", format(round(object$rscore, 2)),
             "  p=", format(1 - pchisq(object$rscore, df)), "\n",
             sep = "")
    invisible(return(list(temp = temp, tmp = tmp)))
  }




N         = 1000                    # number of subjects
fu.min    = 1                       # min follow up time of 1 year
fu.max    = 10                      # max follow up time of 10 years
cens.prob = 0.4                     # censoring probability
dist.x    = "binomial"              # distribution of the covariate
par.x     = list(0.4)         # parameter of the covariate binomial dist
beta      = 0.8                   # regression coefficient
dist.z    = "gamma"                 # distribution of the frailty term
par.z     = 0.7                     # variance θ of the frailty term
dist.rec  = "weibull"               # form of the baseline hazard
par.rec   = c(0.07,1)           # parameters for the distribution of event
pfree     = 1                       # probability of being risk free
dfree     = runif(1,0.5,0.8)  # length of risk free interval

nboot     =   1000                          # number of iterations
modnames  <-
c("cox.I","Marginal.II","PWP.GT.III","frailty.IV","fr.1st.V","Strat.fr.VI","cox.
VII","AG.VIII","PWP.CT.IX","frailty.X","fr.1st.XI","Strat.fr.XII")
coef.m    =   matrix(0,nboot,12,dimnames = list(names(nboot),modnames))
se.m      =   matrix(0,nboot,12,dimnames = list(names(nboot),modnames))
theta.m    =   matrix(0,nboot,12,dimnames = list(names(nboot),modnames))
p.m        =   matrix(0,nboot,12,dimnames = list(names(nboot),modnames))
lb1.m      =   matrix(0,nboot,12,dimnames = list(names(nboot),modnames))
ub1.m      =   matrix(0,nboot,12,dimnames = list(names(nboot),modnames))

my.data1   =   list()


a      <- 1:nboot
b      <- rep(NA,100)
i      <- 1
```

```
warn <- function(w) {
  if(grepl("Inner loop failed to coverge",
           as.character(w))) {aa <<- aa+1}
}

while (i <= nboot){

  b[i] <- a[i]
  aa  <- 0

  my.data1[[i]] <- simrec(N, fu.min, fu.max, cens.prob, dist.x, par.x, beta,
dist.z, par.z,dist.rec, par.rec, pfree, dfree)
  my.data1[[i]]$sort <- ave(my.data1[[i]]$start,my.data1[[i]]$id  ,FUN =
seq_along)
  my.data1[[i]]$tstart<-my.data1[[i]]$start*365.25
  my.data1[[i]]$tstop<-my.data1[[i]]$stop*365.25
  my.data1[[i]]$FUtime<-my.data1[[i]]$tstop-my.data1[[i]]$tstart
  ### Model I - Simple Cox ###
  tryCatch(fitI<-coxph(formula = Surv(FUtime, status) ~ x , data
=my.data1[[i]]), warning = warn)
  coef.m[i,1]   <-fitI$coef[1]
  se.m[i,1]     <-sqrt(fitI$var[1])
  lb1.m[i,1]     <-fitI$coef[1]-qnorm(0.975)*sqrt(fitI$var[1])
  ub1.m[i,1]     <-fitI$coef[1]+qnorm(0.975)*sqrt(fitI$var[1])
  ### Model II - Marginal Cox ###
  tryCatch(fitII<-coxph(formula = Surv(FUtime, status) ~ x +cluster(id), data
=my.data1[[i]]), warning = warn)
  coef.m[i,2]   <-fitII$coef[1]
  se.m[i,2]     <-sqrt(fitII$var[1])
  lb1.m[i,2]     <-fitII$coef[1]-qnorm(0.975)*sqrt(fitII$var[1])
  ub1.m[i,2]     <-fitII$coef[1]+qnorm(0.975)*sqrt(fitII$var[1])
  ### Model III - PWP-GT ###
  tryCatch(fitIII<-coxph(formula = Surv(FUtime, status) ~ x +cluster(id)
+strata(sort), data =my.data1[[i]]), warning = warn)
  coef.m[i,3]   <-fitIII$coef[1]
  se.m[i,3]     <-sqrt(fitIII$var[1])
  lb1.m[i,3]     <-fitIII$coef[1]-qnorm(0.975)*sqrt(fitIII$var[1])
  ub1.m[i,3]     <-fitIII$coef[1]+qnorm(0.975)*sqrt(fitIII$var[1])
  ### Model IV - Frailty ###
  tryCatch(fitIV<-coxph(formula = Surv(FUtime, status) ~ x +frailty(id), data
=my.data1[[i]], method = "em"), warning = warn)
  coef.m[i,4]   <-fitIV$coef[1]
  se.m[i,4]     <-sqrt(fitIV$var[1])
  theta.m[i,4]  <-fitIV$history[[1]]$theta
  lb1.m[i,4]     <-fitIV$coef[1]-qnorm(0.975)*sqrt(fitIV$var[1])
  ub1.m[i,4]     <-fitIV$coef[1]+qnorm(0.975)*sqrt(fitIV$var[1])

  yo.iv<-my.summary.coxph.penal(fitIV)
  p.m[i,4]      <-yo.iv$temp[grepl("frailty",dimnames(yo.iv$temp)[[1]]),"p"]

  ### Model VI - Conditional Frailty ###
  tryCatch(fitVI<-coxph(formula = Surv(FUtime, status) ~ x +frailty(id)
+strata(sort), data =my.data1[[i]],method="em"), warning = warn)
  coef.m[i,6]   <-fitVI$coef[1]
  se.m[i,6]     <-sqrt(fitVI$var[1])
  theta.m[i,6]  <-fitVI$history[[1]]$theta
  lb1.m[i,6]    <-fitVI$coef[1]-qnorm(0.975)*sqrt(fitVI$var[1])
  ub1.m[i,6]    <-fitVI$coef[1]+qnorm(0.975)*sqrt(fitVI$var[1])

  yo.vi          <-my.summary.coxph.penal(fitVI)
  p.m[i,6]      <-yo.vi$temp[grepl("frailty",dimnames(yo.vi$temp)[[1]]),"p"]

  ### Model VII - Simple Cox ###
  tryCatch(fitVII<-coxph(formula = Surv(tstart,tstop, status) ~ x , data
=my.data1[[i]]), warning = warn)
  coef.m[i,7]   <-fitVII$coef[1]
  se.m[i,7]     <-sqrt(fitVII$var[1])
  lb1.m[i,7]    <-fitVII$coef[1]-qnorm(0.975)*sqrt(fitVII$var[1])
  ub1.m[i,7]    <-fitVII$coef[1]+qnorm(0.975)*sqrt(fitVII$var[1])
  ### Model VIII - Marginal Cox ###
  tryCatch(fitVIII<-coxph(formula = Surv(tstart,tstop, status) ~ x +cluster(id),
data =my.data1[[i]]), warning = warn)
  coef.m[i,8]   <-fitVIII$coef[1]
  se.m[i,8]     <-sqrt(fitVIII$var[1])
  lb1.m[i,8]    <-fitVIII$coef[1]-qnorm(0.975)*sqrt(fitVIII$var[1])
  ub1.m[i,8]    <-fitVIII$coef[1]+qnorm(0.975)*sqrt(fitVIII$var[1])
  ### Model IX - PWP-CT ###
  tryCatch(fitIX<-coxph(formula = Surv(tstart,tstop, status) ~ x +cluster(id)
+strata(sort), data =my.data1[[i]]), warning = warn,error=function(e) e)
  cc<-tryCatch(fitIX<-coxph(formula = Surv(tstart,tstop, status) ~ x
+cluster(id)+strata(sort), data =my.data1[[i]]), warning =
warn,error=function(e) e)
  coef.m[i,9]   <-fitIX$coef[1]
  se.m[i,9]     <-sqrt(fitIX$var[1])
  lb1.m[i,9]    <-fitIX$coef[1]-qnorm(0.975)*sqrt(fitIX$var[1])
  ub1.m[i,9]    <-fitIX$coef[1]+qnorm(0.975)*sqrt(fitIX$var[1])
  ### Model X - Frailty ###
  tryCatch(fitX<-coxph(formula = Surv(tstart,tstop, status) ~ x +frailty(id),
data =my.data1[[i]],method="em"), warning = warn)
  coef.m[i,10]  <-fitX$coef[1]
  se.m[i,10]    <-sqrt(fitX$var[1])
  theta.m[i,10] <-fitX$history[[1]]$theta
  lb1.m[i,10]    <-fitX$coef[1]-qnorm(0.975)*sqrt(fitX$var[1])
  ub1.m[i,10]    <-fitX$coef[1]+qnorm(0.975)*sqrt(fitX$var[1])

  yo.x          <-my.summary.coxph.penal(fitX)
  p.m[i,10]      <-yo.x$temp[grepl("frailty",dimnames(yo.x$temp)[[1]]),"p"]


  ### Model XII - Conditional Frailty ###
  tryCatch(fitXII<-coxph(formula = Surv(tstart,tstop, status) ~ x +frailty(id)
+strata(sort), data =my.data1[[i]],method="em"), warning = warn)
  coef.m[i,12]  <-fitXII$coef[1]
  se.m[i,12]    <-sqrt(fitXII$var[1])
  theta.m[i,12] <-fitXII$history[[1]]$theta
  lb1.m[i,12]    <-fitXII$coef[1]-qnorm(0.975)*sqrt(fitXII$var[1])
  ub1.m[i,12]    <-fitXII$coef[1]+qnorm(0.975)*sqrt(fitXII$var[1])

  yo.xii          <-my.summary.coxph.penal(fitXII)
  p.m[i,12]      <-yo.xii$temp[grepl("frailty",dimnames(yo.xii$temp)[[1]]),"p"]



  if (aa==0 & length(cc$message)==0 ){
    i <- i+1

  }else{
    bb<-aa
    print(paste("Repeating i =", i))

  }
}

mean.coef1  = matrix(0,12,1,dimnames=list(modnames))
mean.se1    = matrix(0,12,1,dimnames=list(modnames))
mean.theta  = matrix(0,12,1,dimnames=list(modnames))
cov.m1      = matrix(0,12,1,dimnames=list(modnames))
cov.m3      = matrix(0,12,1,dimnames=list(modnames))

temp=p.m
reject=matrix(0,nboot,12)
for(i in 1:nboot){
  for(j in 1:12){
    reject[i,j]<-as.numeric(temp[i,j])
  }
}

for (i in 1:length(reject))
  if (reject[[i]]<0.05){
    reject[[i]]<-0
  }else{
    reject[[i]]<-1
  }

rej.rate    = matrix(0,12,1,dimnames=list(modnames))



for (i in 1:12){
  mean.coef1[i]  <- mean(coef.m[,i])
  mean.se1[i]    <- mean(se.m[,i])
  mean.theta[i]  <- mean(theta.m[,i])
  cov.m1[i]      <- mean(lb1.m[,i]<=0.8 & ub1.m[,i]>=0.8)
  rej.rate[i]    <- mean(reject[,i])
}
library(haven)
library(frailtypack)
library(foreign)
h<-read.dta("!!gia tin R.dta",convert.dates = TRUE)
save(h, file="h2.rda")
h<-NULL
load("h2.rda")

h2$start <- as.POSIXct(h2$tstart, format="%d/%m/%Y")
h2$start <- as.numeric(h2$start)
h2$stop <- as.POSIXct(h2$tstop, format="%d/%m/%Y")
h2$stop <- as.numeric(h2$stop)

setwd("G:/! DATA/final")

####Gap Timescale#####

### Model I - Cox ###
coxgap<-coxph(formula = Surv(FUtime, gap) ~ age + sex + origin + education +
psoi + Haart + hcvhbv + logvirusload + CD4log , data = h2)
summary(coxgap)
coxgap$loglik

### Model II - Marginal Cox ###
coxgapmarg<-coxph(formula = Surv(FUtime, gap) ~ age + sex + origin + education +
psoi + Haart + hcvhbv + logvirusload + CD4log +cluster(PatientCode) , data = h2)
summary(coxgapmarg)
coxgapmarg$loglik

### Model III - PWP-GT ###
pwpgt<-coxph(formula = Surv(FUtime, gap) ~ age + sex + origin + education + psoi
+ Haart + hcvhbv + logvirusload + CD4log +cluster(PatientCode)+strata(sort),
data = h2)
summary(pwpgt)
pwpgt$loglik

### Model IV - Frailty ###
fgap<-coxph(formula = Surv(FUtime, gap) ~ age + sex + origin + education + psoi
+ Haart + hcvhbv + logvirusload + CD4log +frailty(PatientCode), data = h2,
method = "em")
```

```
summary(fgap)
fgap$loglik


### Model V - Frailty strat ###
fgapstrat<-coxph(formula = Surv(FUtime, gap) ~ age + sex + origin + education +
psoi + Haart + hcvhbv + logvirusload + CD4log +frailty(PatientCode)
+strata(sort), data = h2, method = "em")
summary(fgapstrat)
fgapstrat$loglik

####Calendar Timescale#####

### Model VI - Simple Cox ###
coxcal<-coxph(formula = Surv(start,stop, gap) ~ age + sex + origin + education +
psoi + Haart + hcvhbv + logvirusload + CD4log , data = h2)
summary(coxcal)
coxcal$loglik

### Model VII - AG ###
coxcalcluster<-coxph(formula = Surv(start,stop, gap) ~ age + sex + origin +
education + psoi + Haart + hcvhbv + logvirusload + CD4log
+cluster(PatientCode) , data = h2)
summary(coxcalcluster)
coxcalcluster$loglik

### Model VIII - PWP-GT ###
pwpct<-coxph(formula = Surv(start,stop, gap) ~ age + sex + origin + education +
psoi + Haart + hcvhbv + logvirusload + CD4log +cluster(PatientCode)
+strata(sort), data = h2)
summary(pwpct)
pwpct$loglik


### Model IX - Frailty ###
fcal<-coxph(formula = Surv(start,stop, gap) ~ age + sex + origin + education +
psoi + Haart + hcvhbv + logvirusload + CD4log +frailty(PatientCode), data = h2,
method = "em")
summary(fcal)
```

```
fcal$loglik


### Model XII - Frailty strat ###
fcalstrat<-coxph(formula = Surv(start,stop, gap) ~ age + sex + origin +
education + psoi + Haart + hcvhbv + logvirusload + CD4log +frailty(PatientCode)
+strata(sort), data = h2, method = "em")
summary(fcalstrat)
fcalstrat$loglik


resII<-resid(coxgapmarg,type="martingale",collapse = h2$PatientCode)
resIII<-resid(pwpgt,type="martingale",collapse = h2$PatientCode)
resIV<-resid(fgap,type="martingale",collapse = h2$PatientCode)
resV<-resid(fgapstrat,type="martingale",collapse = h2$PatientCode)
resVII<-resid(coxcalcluster,type="martingale",collapse = h2$PatientCode)
resVIII<-resid(pwpct,type="martingale",collapse = h2$PatientCode)
resIX<-resid(fcal,type="martingale",collapse = h2$PatientCode)
resX<-resid(fcalstrat,type="martingale",collapse = h2$PatientCode)
age<-aggregate(h2$age,by=list(h2$PatientCode), FUN=mean)
par(mfrow=c(4,3))
p1<-plot(h2$age,coxgap$residuals,main = "Model I - Cox",ylab="Martingale
residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p2<-plot(age$x,resII,main = "Model II - Marginal Cox",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p3<-plot(age$x,resIII,main = "Model III - PWP-GT",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p4<-plot(age$x,resIV,main = "Model IV - Frailty",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p5<-plot(age$x,resV,main = "Model V - Conditional Frailty",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p6<-plot(h2$age,coxcal$residuals,main = "Model VI - Cox",ylab="Martingale
residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
```

```
p7<-plot(age$x,resVII,main = "Model VII - AG",sub = "(years)",ylab="Martingale
residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p8<-plot(age$x,resVIII,main = "Model VIII - PWP-CT",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p9<-plot(age$x,resIX,main = "Model IX - Frailty",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
p10<-plot(age$x,resX,main = "Model X - Conditional Frailty",sub =
"(years)",ylab="Martingale residuals",xlab = "Age at initiation of monitoring")
abline(h=0,col="red")
```

## Bibliography

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B.N.; Csáki, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281.

Alter, M. J. (2006). 'Epidemiology of viral hepatitis and HIV co-infection.' *Journal of hepatology*, 44, S6-S9.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012), 'Statistical models based on counting processes', Springer Science & Business Media.

Andersen, P. K., Gill, R. D.(1982), 'Cox's regression model for counting processes: a large sample study*',  Annals of Statistics,* 10, 1100-1120

AVERT(2015), 'What are HIV and AIDS?', URL: http://www.avert.org/about-hiv-aids/what-hiv-aids

Barlow, W. E., & Prentice, R. L.(1988), 'Residuals for relative risk regression*',  Biometrika,* 75(1), 65-74

Breslow, N. (1974), 'Covariance analysis of censored survival data*',  Biometrics*, 89-99

Box-Steffensmeier, J. M., & De Boef, S. (2006). 'Repeated events survival models: the conditional frailty model.' *Statistics in medicine*, 25(20), 3518.

Chi, B.H., Yiannoutsos, C.T., Westfall, A.O., Newman, J.E., Zhou, J., et al. (2011), 'Universal Definition of Loss to Follow-Up in HIV Treatment Programs: A Statistical Analysis of 111 Facilities in Africa, Asia, and Latin America*',  PLoS Med, 8*(10), e1001111

Co-infection: new battlegrounds in HIV/AIDS. Lancet Infect Dis 2013; 13: 559

Clement, D. Y., & Strawderman, R. L. (2009). “Conditional GEE for recurrent event gap times.” *Biostatistics*, *10*(3), 451-467.

Cohen, M.S. et al (2011), 'Prevention of HIV-1 Infection with Early Antiretroviral Therapy*',* *The New England Journal of Medicine,* 367(5), 399-410

Collett, D. (2003), 'Modelling Survival Data in Medical Research, Second Edition', Chapman & Hall/CRC.

Commenges, D., Rondeau, V.(2000), 'Standardized martingale residuals applied to grouped left truncated observations of dementia cases*',* *Lifetime Data Anal,* 6(3), 229-235

Cook, R. J., & Lawless, J. (2007). ‘The statistical analysis of recurrent events.’ Springer Science & Business Media.

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

Duchateau, L., Janssen, P. (2008), 'The Frailty Model', Springer Science & Business Media.

Duchateau, L., Janssen, P., Kezic, I., & Fortpied, C. (2003), 'Evolution of recurrent asthma event rate over time in frailty models*',* *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 52(3), 355-363

ECDC (2015), 'HIV/AIDS surveillance in Europe 2014', URL: http://ecdc.europa.eu/en/publications/Publications/hiv-aids-surveillance-in-Europe-2014.pdf

ECDC (2015), 'Highest number of new HIV cases in Europe ever', URL: http://ecdc.europa.eu/en/press/Press%20Releases/highest-number-new-HIV-cases-ever-26-November-2015.pdf

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). 'Applied longitudinal analysis' (Vol. 998). John Wiley & Sons.

Fleming, T. R., & Harrington, D. P. (2011), 'Counting processes and survival analysis', John Wiley & Sons.

Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W.J., Burke, G., McBurnie, M.A. (2001), 'Frailty in older adults: Evidence for a phenotype*', Journals of Gerontology - Series A Biological Sciences and Medical Sciences,* 56(3), M146-M156

Gill, R. D.(1984), 'Understanding cox's regression model: a martingale approach*', Journal of the American Statistical Association,* 79(386), 441–447

Gray, R. J. (1992). 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, *87*(420), 942-951.

Grenander, U., (1981), 'Abstract inference', John Wiley & Sons.

Hanagal, D. D. (2011), 'Modelling survival data using frailty models', CRC Press.

Henderson, R., & Oman, P. (1999). "Effect of frailty on marginal regression estimates in survival analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,61(2), 367-379.

Hosmer, D.W., Lemeshow, S., May, S. (2008), 'Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd Edition', John Wiley & Sons.

Hougaard, P. (2000), 'Analysis of Multivariate Survival Data', Springer.

Hougaard, P., Harvald, B., Holm, N. V. (1992), 'Measuring the Similarities Between the Lifetimes of Adult Danish Twins Born Between 1881-1930*', Journal of the American*

*Statistical Association,* 87(417), 17-24

Huang, C. Y., Qin, J., & Wang, M. C. (2010), 'Semiparametric Analysis for Recurrent Event Data with Time-dependent Covariates and Informative Censoring', *Biometrics,* 66(1), 39-49

Kalbfleisch, J.D., Prentice, R.L. (2002), 'The Statistical Analysis of Failure Time Data', John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958), 'Nonparametric Estimation from Incomplete Observations', *Journal of the American Statistical Association,* 53, 457-481

KEELPNO (2015), 'HIV infection: New epidimiologic data, October 2015 '

KEELPNO (2014), 'HIV/AIDS Surveillance in Greece Data reported through 31.12.2014'

Kelly, P. J., & Lim, L. L. Y.(2000), 'Survival analysis for recurrent event data: an application to childhood infectious diseases', *Statistics in medicine,* 19(1), 13-33

Klein, J. P., Moeschberger, M. L. (2003), 'Survival analysis Techniques for censored and truncated data', Springer-Verlag.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (2013), 'Handbook of survival analysis', CRC Press.

Kleinbaum, D. G., Klein, M. (2006), 'Survival analysis: a self-learning text', Springer Science & Business Media.

Lee, E. W., Wei, L. J. and Amato, D. A(1992), 'Cox-type regression analysis for large numbers of small groups of correlated failure time observations', in Klein, J. P. and Goel, P. K. (eds), Survival Analysis: State of the Art, Kluwer Academic Publisher, Dordrecht, 237-247

Lim, H. J., Liu, J., & Melzer-Lange, M. (2007). “Comparison of methods for analyzing recurrent events data: application to the emergency department visits of pediatric firearm victims.” *Accident Analysis & Prevention*, 39(2), 290-299.

Lin, D. Y., Wei, L. J. (1989). ‘The robust inference for the Cox proportional hazards model.’, *Journal of the American Statistical Association,* 84: 1074–1078

Malaza, A., Mossong, J., Bärnighausen, T., Viljoen, J., & Newell, M. L. (2013). 'Population-based CD4 counts in a rural area in South Africa with high HIV prevalence and high antiretroviral treatment coverage.', *PloS one*, 8(7), e70126.

Mazroui, Y., Mauguen, A., Mathoulin-Pélissier, S., MacGrogan, G., Brouste, V., & Rondeau, V. (2016), 'Time-varying coefficients in a multivariate frailty model: Application to breast cancer recurrences of several types and death*',* *Lifetime data analysis,* 22(2), 191-215

Murphy, S. A., & Sen, P. K. (1991), 'Time-dependent coefficients in a Cox-type regression model', *Stochastic Processes and their Applications*, 39(1), 153-180.

National Institute of Allergy and Infectious Diseases (2015), 'How HIV causes AIDS', URL: http://www.niaid.nih.gov/topics/hivaids/understanding/howhivcausesaids/Pages/howhiv.aspx

Platt, L., Easterbrook, P., Gower, E., McDonald, B., Sabin, K., McGowan, C., ... & Vickerman, P. (2016), ‘Prevalence and burden of HCV co-infection in people living with HIV: a global systematic review and meta-analysis’ *The Lancet infectious diseases*, 16(7), 797-808.

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981), 'On the regression analysis of multivariate failure time data*',* *Biometrika*, 68, 373-379

Quinn, T. C., Wawer, M. J., Sewankambo, N., Serwadda, D., Li, C., Wabwire-Mangen, F., ... & Gray, R. H. (2000). 'Viral load and heterosexual transmission of human immunodeficiency virus type 1', *New England journal of medicine*, 342(13), 921-929.

Rockwood, K., Stadnyk, K., MacKnight, C., McDowell, I., Hébert, R., Hogan, D.B.(1999), 'A brief clinical instrument to classify frailty in elderly people', *Lancet*, 353 (9148), 205-206

Rondeau, V., Commenges, D., & Joly, P.(2003), 'Maximum penalized likelihood estimation in a gamma-frailty model', *Lifetime data analysis,* 9(2), 139-153

Therneau, T. M., & Grambsch, P. M. (2000), 'Modeling survival data: extending the Cox model', Springer Science & Business Media.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R.(1990), 'Martingale-based residuals for survival models', *Biometrika,* 77(1), 147-160

Therneau, T. M., & Lumley, T. (2016). Package 'survival'.
URL: https://CRAN.R-project.org/package=survival

Thierfelder, C., Weber, R., Elzi, L., Furrer, H., Cavassini, M., Calmy, A., ... & Ledergerber, B. (2012). "Participation, characteristics and retention rates of HIV-positive immigrants in the Swiss HIV Cohort Study." *HIV medicine*, 13(2), 118-126.

Ullah, S., Gabbett, T. J., & Finch, C. F. (2012). "Statistical modelling for recurrent events: an application to sports injuries." *British journal of sports medicine*, bjsports-2011.

UNAIDS (2014), Global AIDS epidemic facts and figures. URL:http://www.unaids.org/sites/default/files/media_asset/20140716_FactSheet_en.pdf

UNAIDS (2014), 'Fast-Track - Ending the AIDS epidemic by 2030', URL: http://www.unaids.org/sites/default/files/media_asset/JC2686_WAD2014report_en.pdf

UNAIDS (2015), 'AIDS by the numbers 2015', URL: http://www.unaids.org/sites/default/files/media_asset/AIDS_by_the_numbers_2015_en.pdf

Van Beckhoven, D., Florence, E., Ruelle, J., Deblonde, J., Verhofstede, C., Callens, S., For the BREACH (Belgian Research on AIDS and HIV Consortium). (2015). 'Good continuum of HIV care in Belgium despite weaknesses in retention and linkage to care among migrants.' *BMC Infectious Diseases*, 15, 496

Vaupel, J.W., Manton, K.G., Stallard, E. (1979), 'The impact of heterogeneity in individual frailty on the dynamics of mortality*', Demography*,16(3), 439-454

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989), 'Regression analysis of multivariate incomplete failure time databy modeling marginal distributions*', Journal of the American Statistical Association,* 84, 1065-1073

Wienke, A. (2010), 'Frailty models in survival analysis', CRC Press.

World Health Organisation (2013), 'Consolidated Guidelines On The Use Of Antiretroviral Drugs For Treating And Preventing HIV Infection',

URL: http://apps.who.int/iris/bitstream/10665/85321/1/9789241505727_eng.pdf

World Health Organisation(2002), 'Scaling Up Antiretroviral Therapy In Resource-Limited Settings (Guidelines For A Public Health Approach)',

URL: http://www.who.int/hiv/pub/prev_care/ScalingUp_E.pdf

World Health Organisation (2015), 'Guideline On When To Start Antiretroviral Therapy And On Pre-exposure Prophylaxis For HIV'

Yu, L. M., Easterbrook, P. J., & Marshall, T. (1997). 'Relationship between CD4 count and CD4% in HIV-infected people', *International journal of epidemiology*, 26(6), 1367-1372.

Yu, Z., Liu, L., Bravata, D. M., Williams, L. S., & Tepper, R. S. (2013). 'A Semiparametric recurrent events events model with time-varying coefficients', *Statistics in medicine,* 32(6), 1016-1