Defaulters in a cohort of HIV infected patients

Oluwaseyi Akindunjoye

promotor :
De heer Pryseley ASSAM NKOUIBERT, Dr.
Eric FLORENCE

Universiteit Maastricht

Eindverhandeling voorgedragen tot het bekomen van de graad
Master of science in Applied Statistics

universiteit hasselt

Universiteit Hasselt
Center for Statistics

Defaulters in a cohort of HIV infected patients.

By

Oluwaseyi, Samson Akindunjoye

Assam, Pryseley — **Internal Supervisor**

Dr. Eric, Florence — **External Supervisor**

Thesis submitted to the Center for Statistics, Universiteit Hasselt, Diepenbeek, Belgium, in partial fulfillment of the requirement for the award of Master of Science in Applied Statistics.

September 2007

Certification

This is to certify that this research work was carried out by **Oluwaseyi, Samson Akindunjoye** under our supervisions.

Oluwaseyi, Samson Akindunjoye

…….

Student

Assam, Pryseley	**Dr. Eric, Florence**
……	**……**
Internal Supervisor	***External Supervisor***

Dedication

To my mother

and

the memory of my father

Abstract

The advent of antiretroviral therapy (ART) has transformed HIV/AIDS from a primary deadly disease into a chronic disease characterized by enhanced quality of life and increased life expectancy. Once started, the antiretroviral treatment should be continued lifelong and adherence to this treatment should be nearly perfect to enable long-term efficacy. Despite the improvements in management in Europe, HIV infected persons still remain vulnerable to drop out/loss to follow up from care and treatment. Therefore, this research tries to review the defaulter rate during the last five years (2002-2006) at the HIV outpatient clinic and how it is evolving during this period.

Data was explored using Kaplan-Meier curve to know whether or not the groups are proportional through the assumption of proportional hazards i.e. if the estimated survival functions for two groups of survival data are approximately parallel (do not cross) and a Logrank and Gehan-Wilcoxon tests were used to compare the survival estimates between two or more groups. The Survival data is modeled using Cox's proportional Hazard model to explore the relationship between survival and explanatory variables thereby analyzing for the effect of several risk factors on survival.

Collett's approach criterion was applied to select the best model ignoring the missingness mechanism in the data. In order to assess the adequacy of the fitted model, residual plots and some formal tests (time-dependent covariate) were used to check for the assumption of proportional hazard. A stratified analysis was carried out to compare the Cox's proportional hazard model and the stratified model which tells us if the fitted model is good or not. The use of single and multiple imputation methods were used to investigate the nature of the missingness mechanism and its effect, observing the possibility of Missing at Random (MAR).

Based on this study, there was an association between defaulters and gender, risk group, clinical stage, sex preference, origin group, ART, viral load and age group, also an evolution over time shows how the patients default through an increasing trend. Our analysis showed that 13.97% of 1167 patients defaulted while 8.94% were lost to follow-up but out of 163 defaulters, more than half of them defaulted in 2006 alone (53.99%) while 3.68% defaulted in the first year of ART. Of all the patients, 67.15% were male but in proportion, female defaulted more than male with 19.84% and 13.05% of female were lost to follow-up. It was observed that the model does not satisfy the proportional hazard assumption even with time dependent covariates, the global goodness of fit test shows that the model is fitted at a borderline significant level but correcting for missingness, the PH assumptions hold in both responses.

Keywords: Antiretroviral Therapy (ART), Kaplan-Meier Plot (KM), Cox's Proportional Hazard Model (PH)

Table of Contents

Contents **Page**

List of Tables

List of Figures

Acronym

AIDS	= Acquired Immune Deficiency Syndrome
ART	= Antiretroviral Therapy
AVGCD4	=Average CD4
AVGVL	= Average Viral Load
CD4	= T-cells or T-lymphocytes (White blood cell)
HAART	=Highly Active Antiretroviral Therapy
HIV	= Human Immunodeficiency Virus
KM	=Kaplan-Meier
PH	=Proportional Hazard
VL	= Viral Load
HR	= Hazard Ratio
CI	= Confidence Interval
CMH	= Cochran-Mantel Haenzel
OR	= Odds Ratio

Acknowledgements

First of all, I would like to give thanks to Almighty God who has given me the opportunity to come and study in the University Hasselt.

My unreserved appreciations also go to my internal supervisor who shared with me his long-term experience and never minded disturbing him whenever I was at a cross road regarding the way forward in the process of analyzing the dataset. In addition, I wish to say a big thank to my external supervisor for sharing their invaluable knowledge with me, provision of useful materials to ease my task and always attending to my requests and questions.

Knowledge gained from the individual and group interactions contributed a lot to improve this Thesis and as a result of this I thank my fellow course mates in Applied Statistics for their support and enthusiasm.
I am also indebted to the following senior colleagues for their advice and assistance rendered when needed; daddy Aishat, Tayo, Lamidi, Hermann, Chiara, Susan.
A special thanks goes to all my friends for their contribution to my success in Belgium: Segun, Ayo, Funmi, Nurudeen, Titi, Doyin, Dapo, Ebun, Nzume and Kunle Badru. It will be unfair if I don't appreciate my neighbours who made life more interesting for me in Belgium through their simple way of life.

I thank my caring mother for her prayer, advice and her determination to assist me in pursuing further studies outside my country. To all my siblings, niece and nephews, you are all cherished for your love and unity in the family.
I have to commend the persistent, patience, perseverance, understanding and support of my fiancé, Seun Ojo. Her everyday sacrifice on calls motivates me to achieving my targeting goal in the course of my study. You will surely be rewarded with a unique gift.

Finally, I acknowledge the assistant receive from Prof. Florin Vaida during and after the Biostatistics survival class and those people that I used their materials in the cause of this research work.

1. INTRODUCTION

1.1 Background Knowledge

HIV (Human Immunodeficiency Virus) is a virus that causes AIDS. HIV attacks the body's immune system and destroys certain blood cells e.g the CD4 cells that are crucial to the normal function of the immune system, which defends the body against illness (diseases and tumors). Months to years after a person is infected with HIV, the virus destroys all the infection fighting cells called T-cell lymphocytes and makes the host susceptible to opportunistic infections which cause severe or fatal health problems and could cause the death of the HIV patient. However, HIV can also attack cells of the brain, nervous system, digestive system, lymphatic system, and other parts of the body and also depends on some factors like treatment duration, infections the person is exposed to [18]. When a person with an HIV-weakened immune system has a CD4 cell count below 200 or 14%, that person may be diagnosed by a doctor as having AIDS. HIV is the primary agent that leads to the development of AIDS. A person without drug treatment could still depend on the reasonable diet taken and being malnourished if not HIV patients can live for a long time before it becomes AIDS [19].

An HIV positive test result means that a person has been infected and could transmit it to others. The test does not look for the actual virus itself, but found evidence of it in the blood. Though, very difficult to tell from this result who transmitted the virus, for how long or when it will begin to affect the health.

HIV is spread through these routes:

- Transfusion of blood and blood products - this includes via blood transfusion (rare in the UK).
- Use of infected donor organs, tissue or semen.
- Mother to baby - 17% - 30% of babies born to HIV +ve mother will themselves become infected which could only be established after a certain time, also during pregnancy and through breast feeding of infected mother.
- Contamination with infected blood like sharing of needles for intravenous drug use e.t.c.
- Unprotected penetrative vaginal or anal intercourse - heterosexual sex and also homosexual males. In the UK, 60% of HIV diagnoses are from homosexual or bisexual men. Heterosexual intercourse accounts for 19% of HIV.

The alarming trend on HIV is that everyday about 14,000 new HIV infections occur worldwide. Approximately 33.3 million people in the world are living with HIV. Of these, more than 95% are in low and middle income countries, almost 2000 are among children under 15,over 40% occur among women and more than 40% are among young people aged 15 to 24.Globally, of all the people living with HIV, less than half are female. Over one-third (36%) of people living with HIV in Latin America are female. Women and girls represent 57% of all the people. Also, in Sub-Saharan Africa, a striking 76% of young people (aged 15-24yrs) livings with HIV are female. As of June 2005, an estimated 6.4million people in low and middle-income countries were in need of antiretroviral treatment, yet only 1million (15% were receiving such care) [15]. With 290,000 or 62% of those in need of HIV treatment receiving care, the Latin

America/Caribbean region has the highest rate of coverage but in Sub-Saharan Africa, an estimated 11% of the 4.7million people in need of HIV treatment are receiving care [16].

The antiretroviral therapy (ART) is a drug type that handles HIV infection, it slows down the reproduction and the progression of HIV disease thereby enhancing good quality of life and increased life expectancy. ART can prolong the time between HIV infection and the onset of AIDS. This treatment, though not a cure, enables people living with HIV to enjoy longer, healthier lives, and as such it acts as an incentive for people to volunteer for HIV testing. The treatment consists of drugs that have to be taken every day for the rest of someone's life. These medicines, however, are not widely available in many poor countries around the world, and millions of people who cannot access medication continue to die.ART supposed to be a continual lifelong treatment to enable a long-term efficacy but still HIV infected persons remain vulnerable to drop out/loss to follow up from care and treatment defaulting from treatment even despite the improvement in management in Europe.

More than one antiretroviral drug at a time is necessary for ART to be effective for a long time. The term Highly Active Antiretroviral Therapy (HAART) is used to describe a combination of three or more anti-HIV drugs known as combination Therapy. When HIV replicates (makes new copies of itself) it often makes mistakes. This means that within any infected person there are many different strains of virus. Occasionally, a new strain is produced that happens to be resistant to the effects of an antiretroviral drug. If the person is not taking any other type of drug then the resistant strain is able to replicate quickly and the benefits of treatment are lost. Taking two or more antiretrovirals at the same time vastly reduces the rate at which resistance develops.

1.2 Research Questions

The primary objective of the study is to review the defaulter rate during the last five years (2002-2006) at an HIV outpatient clinic and how it is evolving during this period.

Secondary objectives are to find out the causes of patient default (death, followed in another hospital, loss to follow-up)? What are the risk factors associated with patient default? And which interventions can we set up to decrease the loss to follow-up or get patients back in the medical system?

The motivation behind this study is to know who are defaulters among the patients, rate at which they defaulted and reasons why they defaulted having known that active patients are patients with at least one contact with a physician at Institute of Tropical Medicine, Antwerp, calendar year (1 January till 31 December).

The next section describes the dataset used in this study, section 3 deals with methodology used in the study; section 4 is all about results of the analysis. Discussion of the entire results was given in section 5 and brief conclusion was elucidated in section 6 while some recommendations were given to cap it up in section 7.

The software used in this study are S-Plus 6.2 Professional, SAS version 9.1.3 and R 2.5.0 for graphical presentation and statistical analysis of this study.

2. DATA DESCRIPTION

The source of data is from HIV outpatient clinic, Prince Leopold Institute of Tropical Medicine, Antwerp. The data was extracted from the HIV cohort database. This cohort contains epidemiological, laboratory and clinical information of all HIV patients in follow-up since 2000 - 2006, including a detailed antiretroviral therapy history. Five different data sets which are patient, treatment, consultation, CD4count, viral load datasets were merged together and was used for this analysis. The five datasets are of different observations and were critically managed to form a complete dataset.

Patients dataset contains 5295 subjects with some covariates like the gender of the patients, date of birth, date of first HIV test, origin, clinical stage, sex reference and risk group. Then, variable age at entry was created (date of birth – first consultation date) which was categorized during data exploration. Origins of the patients was categorized according to their continents: Europe, Asia, Africa, America and Oceania. Sex Preference is the sex preferred by the patients and was categorized into heterosexual, homosexual and unknown.

Clinical Stages are classified into three; A, B and C [17]. Clinical Stage A is called Clinically *Asymptomatic* Stage. At this stage, patients are free from major symptoms, although there may be swollen glands. The level of HIV in the peripheral blood drops to very low levels but people remain infectious and HIV antibodies are detectable in the blood, so antibody tests will show a positive result. Research has shown that HIV is not dormant during this stage, but is very active in the lymph nodes. Viral load test is used to measure the small amount of HIV that escapes the lymph nodes. At *Symptomatic* stage, over time the immune system becomes severely damaged by HIV affecting the lymph nodes and tissues leading to more T helper cell destruction. As the immune system deteriorates the symptoms worsen. This stage is mainly caused by the emergence of opportunistic infections and cancers that the immune system would normally prevent. Clinical Stage C is *Aids*, this occurs since as the immune system becomes more and more damaged the illnesses that occur become more and more severe leading eventually to an AIDS diagnosis.

Risk group of HIV patients were into these categories; heterosexual contacts, homosexual contacts, IV drug user, other, unknown. Homosexual contacts are men who have sex with men, heterosexual contacts involve both sex. The IV drug user risk group contact HIV through injecting drug use while unknown risk group are the blood and blood factor recipients, children born to HIV infected mothers e.t.c. Age was also categorized into clinically meaningful groups.

Treatment dataset was repeatedly measured over time having 3850 observations but the patients that went for treatment were 1303. Patients were given different types of treatment and the covariates are the treatment dates (first and last), number of times each patient was treated and the number of treatments given to patients. The outcome variables were created from this dataset, they are defaulters and loss to follow up but focused was more on the defaulters based on the research questions though loss to follow up was also analyzed.

Defaulters are patients that do not come back at least one year after their last medical contact till the data was censored (31 December 2006) or patients that are away for more than one year at anytime during the treatment period. Meanwhile, all patients who came in the last 12 months before 31 December 2006 will be considered as in follow-up. Defaulter could be as a result of

death, transferred in another hospital and loss to follow-up**.** Patients that defaulted are not active patients since they missed contact at least one year and active patients do not default, they have at least one contact with the physician in a calendar year. Loss to follow up are patients who did not come in the last 12 months before 31 December 2006 i.e almost over a year before the data is censored. Any patient that is loss to follow up is a defaulter but a defaulter might not be lost to follow up.

Consultation dataset has 38050 observations and was repeatedly measured over time but the patients that appeared for consultation were 15195. More covariates were created like number of physician during follow up, number of consultation during follow up, number of times each patient was repeatedly measured, first consultation date and last consultation date. Also, a covariate "On antiretroviral treatment at last consultation" is created and categorized as: ART = Yes if the date of last treatment is more than the date of last consultation and No otherwise. ART at last consultation is considered to know how many patients were on treatment when they came last for consultation because once they default, they might soon fall short of drugs and be sick again.

CD4count dataset was measured repeatedly over time having 25882 observations. The covariates were CD4count, CD4 dates. Some other covariates were created like the lowest CD4count in each patient (Nadir CD4 count), date of first CD4count, date of last CD4count, last CD4count, first CD4count, the average CD4 count for each patient which was created by dividing the change in CD4dates (last date - first date) and the change in CD4 count (last CD4 count – first CD4 count). Also another covariate was created from Nadir CD4count to see if patients who were very sick are more prone to default. If Nadir CD4 count is less than 0.2, it is "below CD4count and above CD4count otherwise. It is worthy to know that patients with CD4count < 0.2 could reflect a risk of opportunistic infections, only 2002 patients were measured for CD4count.

Viral load dataset was measured repeatedly over time and the observation was 28360.The covariates were viral load, viral load dates. Some other covariates were created like the date of first viral load, date of last viral load, last viral load, first viral load, the average viral load for each patient which was created by dividing the change in viral load dates (last date - first date) and the change in viral load (last viral load – first viral load), also created is a categorized viral load to know the treatment failure at last consultation (if viral load > 400 copies/ml at last consultation among treated patients, it is high viral load otherwise it is low viral load).This is important since high Viral load means higher HIV disease indicating that the disease is really active and for patients between 200-500 Viral load, their HIV can not reproduce thereby reduces disease progression. If patients are lost with a high viral load, HIV could be transmitted easily and treatment failure associated with default might be that the patients are failing because they are not taking their drugs good as prescribed, only 2281 patients were measured for viral load.

All the five datasets were merged together *by* patient identification number *(patid)*. The whole dataset (merged) has 1167 subjects that were finally used for the analysis. The response variable is the defaulters while the covariates are age at entry in medical care, gender (man, woman, unknown), origin group, risk group, sexual preference, number of consultation during follow up, number of physician during follow-up, ART at last consultation (yes, no, unknown), Nadir CD4

count (lowest CD4), Average CD4 count, Average viral load, On ART at last consultation, CD4 count (below or above) and Viral Load (high or low). All the categorized covariates were done based on expert opinion, the protocols of this research and their clinical importance. There is missingness in the dataset for some of the covariates. The percentage of missing observations varies from one covariate to the other in the complete dataset. Having used a complete-case analysis, where observations with any missing values are excluded from the analysis, knowing that no special analysis is required in it and it discards a lot of information, especially when there is a substantial amount of missing data, losing statistical power. To avoid biased interference result, missingness is checked using Multiple Imputation.

3. STATISTICAL METHODOLOGY

3.1 Exploratory Data Analysis

Exploratory analysis was done in order to gain more insight into the dataset. The statistical tools employed in this section are summary statistics such as mean, standard deviation, minimum and maximum. Graphical illustrations like plots, Multiple-bar charts were used mainly for the categorical variables to see the relationship between defaulters and each covariate.

3.1.1 Kaplan-Meier curve

Kaplan-Meier curve is the most common method to describe survival characteristics which was used for all the categorical predictors to provide insight into the shape of the survival function for each group and give an idea of whether or not the groups are proportional i.e. if the estimated survival functions for two groups of survival data are approximately parallel (do not cross), the assumption of proportional hazards is justified. Kaplan-Meier curves work best for time fixed covariates with few levels. It is not feasible to plot a Kaplan-Meier curve for continuous variable since there will be a curve for each level of the predictor and a continuous predictor has many different levels then the graph becomes too "cluttered". It is worthy to know that assumption of proportional hazards could be checked also through some formal tests to see if the estimated survival curves are parallel. The population will be first described per year with the number of active patients and defaulter per year. Having explored the dataset, defaulter rates will then be estimated using survival analysis methods. Risk factors for defaulters from care and treatment will be assessed using Cox-proportional hazards models. Also, a univariate test with a single continuous predictor was done using a univariate Cox proportional hazard regression (Semi-parametric model).

3.1.2 Non-Parametric Test

In comparing the survival estimates between two groups, a non- parametric approach is used**.** There are different sample tests used in survival data analysis like logrank test (Cochran-Mantel-Haenszel and Peto-Peto / Linear rank test) and Gehan-Wilcoxon test (Peto-Peto-Prentice-Gehan-Wilcoxon test) but focus will be on the Cochran-Mantel-Haenszel type of logrank test and Gehan-Wilcoxon test. Though both tests could lack power if the survival curves (or hazard) "cross" but this does not necessarily make them invalid. The logrank is most powerful under the assumption of proportional hazards,

$$\lambda_1(t) = \alpha\,\lambda_2(t),$$

where α is a constant that does not depend on time, which implies an alternative in terms of the survival functions of Ha : $S1(t) = [S2(t)]^{\alpha}$ .The null hypothesis of no difference between the groups of survival times is that the hazard of default at any given time for an individual in one group is proportional to the hazard at that time for a similar individual group. This is the motivation for interpreting the log rank test result in this research work.

(a) Logrank Test

The logrank test (Cox-Mantel test) is obtained by constructing a (2*2) table at each distinct default time, and comparing the default rates between the two groups, conditional on the number at risk in the groups.
The logrank test is:

$$\chi^2_{\log rank} = \frac{\left[\sum_{j=1}^{k}\left(d_{0j} - r_{0j} * d_j / r_j\right)\right]}{\sum_{j=1}^{k} \frac{r_{1j} r_{0j} d_j \left(r_j - d_j\right)}{\left[\left(r^2{}_j \left(r_j - 1\right)\right)\right]}}$$

where d_{0j} and d_{1j} are the number of deaths in group 0 and 1, respectively at the *j*-th death time, and r_{0j} and r_{1j} are the number at risk at that time, in groups 0 and 1. Assuming the tables are all independent, then this statistic will have an approximate χ^2 distribution with 1 degree of freedom. We motivated the logrank test through the CMH statistic for testing H_0*: OR = 1 over K tables, where K is the number of distinct default times. (*Hosmer and Lemeshow 1

998).

The logrank test is most powerful for proportional hazards when “odds ratios" are constant over time intervals and is used to test the hypothesis of equality of two survivor functions. The logrank test places more weight on larger survival times unlike the Gehan-Wilcoxon test, which places more weight on early survival times. The Gehan-Wilcoxon test is sensitive to early differences in survival btw groups; note $w_j = s(t_j)$ while the log-rank test is sensitive to later differences; note $w_j = 1$.

Set Hypothesis;

Test of equality over strata

H_0: To know if there is no significant difference between the defaulters and each covariate.

H_1: To know if there is significant difference between the defaulters and each covariate

The larger the value of this statistic, the greater the evidence against the null hypothesis, reject the null hypothesis if the p value is less than the 5% level of significance.

(b) Gehan-Wilcoxon (Generalized Wilcoxon) Test

$$Let\ \ U\left(X_i, Y_j\right) = U_{ij} = \begin{cases} +1 \text{ if } x_i > y_j \text{ or } x_i^+ \geq \mathrm{y}_j \\ 0 \quad \text{if } x_i = \mathrm{y}_i \text{ or lower value censored} \\ -1 \text{ if } x_i < y_j \text{or } x_i \leq y_j^+ \end{cases}$$

Then define

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$$

Thus, there is a contribution to W for every comparison where both observations are failures (except for ties), or where a censored observation is greater than or equal to a failure. First, pool the sample of (n+*m)* observations into a single group, then compare each individual with the remaining n+m-1. For comparing the i-th individual with the j-th, define

$$U_{ij} = \begin{cases} +1 \text{ if } \mathrm{t}_i > t_j \text{ or } \mathrm{t}_i^+ \geq t_j \\ \text{-1} \quad \text{if } \mathrm{t}_i < \mathrm{t}_j \text{ or } \mathrm{t}_i \leq t_j^+ \\ 0 \quad \text{if} \quad \textit{otherwise} \end{cases}$$

Then

$$U_i = \sum_{j=1}^{m+n} U_{ij}$$

Thus, for the i-th individual, U_i is the number of observations which are definitely less than t_i minus the number of observations that are definitely greater than t_i. Here, we assume censorings occur after deaths.

The Gehan statistic is defined as

$$U = \sum_{i=1}^{m+n} U_i 1_{\{i \text{ in group } 0\}} \quad = \quad W$$

U has mean 0 and variance

$$\mathrm{var}(u) = \frac{mn}{(m+n)(m+n-1)} \sum_{i=1}^{m+n} U_i^2$$

***(c) P*-Sample Logrank**

This is used when we are comparing survival distributions between more than two groups to know whether the groups differ from each other. Suppose we observe data from P different groups, and the data from group p (p = 1,….., P) are:

$$\left(X_{p1,}\delta_{p1}\right)...\left(X_{pn_p},\delta_{pn_p}\right)$$

Constructing a $(P*2)$ table at each of the *K* distinct default times, and compare the default rates between the *P* groups, conditional on the number at risk. Using the CMH approach then a $\chi^2_{(P-1)}$ test statistic could be constructed through a comparison of "o"s and "e"s, like before.

3.2 Cox's Proportional Hazard Model

Cox's Proportional Hazard regression model is used to explore the relationship between survival and explanatory variables by modeling the survival data thereby analyzing the effect of several risk factors on survival. Cox PH model is a semi-parametric PH models since there is no assumptions concerning the nature or shape of the underlying survival distribution.

Proportional Hazard (PH) Model

$$\lambda(t;\mathbf{Z}) = \lambda_0(t)\ \Psi(Z)$$
$$\lambda(t;Z) = \lambda_0(t)\exp(\beta Z)$$
$$= \lambda_0(t)\exp\left(\sum \beta_j Z_j\right)$$

β_j is the parameter for the j-th covariate (Z_j). $\lambda_0(t)$ is called the *baseline hazard*; it is the hazard for the respective individual when all independent variable values $Z_1,...,Z_p$ are equal to zero (i.e., the "reference group") but in comparing two groups, $\lambda_0(t)$ is the hazard rate for one of the two groups. $\lambda(t;\mathbf{Z})$ denotes the resultant hazard, given the values of the *p* covariates for the respective case ($Z_1,...,Z_p$) and the respective survival time (*t*).

Important Assumptions of this model:

(A1) The baseline hazard $\lambda_0(t)$ depends on *t*, but not on covariates $Z_1,...,Z_p$.

(A2) The hazard ratio, i.e. exp (**βZ**), depends on the covariates $\mathbf{Z} = (Z_1,...,Z_p)$, but not on time *t*.

(A3) In addition, the covariates Z_j do not depend on time *t*.

Hazard Ratio (Φ)

$$\frac{\lambda(t;Z_i)}{\lambda(t;Z_i')} = \frac{\lambda_0(t)\exp(\beta Z_i)}{\lambda_0(t)\exp(\beta Z_i')} = \frac{\exp(\beta Z_i)}{\exp(\beta Z_i')} = \exp\left[\left(\beta\left(Z_{ij} - Z_{ij}'\right)\right)\right] = \Phi$$

In the last formula, Z_{ij} is the value of the *j*-th covariate for the *i*-th individual. For example, Z_{42} might be the value of gender (0 or 1) for the 4-th person. Note that $\lambda(t;Z_i) / \lambda(t;Z_i')$ is constant in time (does not depend on time). If the PH assumption does not hold, then for some Z_i , Z_i', the function $\lambda(t;Z_i) = \lambda(t;Z_i')$ depends on time *t*. $\lambda(t;Z_i)$ and $\lambda(t;Z_i')$ are the hazards of death at time t for patient on one group and patient on the other group.

Hypothesis Tests:

For each covariate of interest, the null hypothesis is

$$H_0 : HR_j = 1,\ \beta_j = 0$$

A Wald test of this hypothesis is used and this test for $\beta_j = 0$ assumes that all other terms in the model are held fixed.

In general, Z is a *vector* of covariates of interest. Z may include continuous, discrete factors and possible interactions. If we have a discrete covariate *A* with *a* levels, then we will need to include (*a*-1) dummy variables ($U_1, U_2 ..., U_a$) such that $U_j = 1$ if $A = j$. Then

$$\lambda_i(t) = \lambda_0(t) \exp\left(\beta_2 U_2 + \beta_3 U_3 + ... + \beta_a U_a\right)$$

(In the above model, the subgroup with $A = 1$ or $U_1 = 1$ is the reference group.) Two factors, *A* and *B*, interact if the hazard of death depends on the combination of levels of *A* and *B*. The principle of hierarchical models are considered interactions are only included if all of the corresponding main effects are also included.

Assumption (A2) above implies that the ratio of the hazards for two individuals with covariates Z_1 and $Z_1^{'}$ is a constant, Φ, which does NOT depend on time, *t*. In other words, the hazards of the two groups remain proportional over time. Φ is the ratio of the hazard of death at any time for an individual in one group relative to an individual in the other group. It is worthy to know that Cox's PH model has the advantage over a simple log-rank test of giving us an estimate of the hazard ratio" (i.e., $\Phi = \lambda_1(t) / \lambda_0(t)$). This is more informative than just a test statistic. **PROC LIFETEST** in SAS is used with **STRATA** option for categorical variables in which we were able to get the estimates of the entire survival distribution $\hat{S}(t)$ for each group. Exact method is applied to correct for ties.

3.3 Model Selection

3.3.1 Model Selection Criteria

3.3.1.1 Likelihood Ratio Test

Suppose there are $(p + q)$ explanatory variables measured and proportional hazards are assumed. Using *nested* models with d.f = *p* and $p + q$. For such nested models, we can construct a likelihood ratio test of $H_0: \beta_{p+1} = ... = \beta_{p+q} = 0$ as:

$$\chi^2_{LR} = -2[\log \text{lik1} - \log \text{lik2}]$$

Select smaller model if χ^2_{LR} *not* significant against χ^2_q distribution i.e. if $\chi^2_{LR} < \chi^2_q$ accept Ho and use smaller model, otherwise. Also, larger difference between -2 log lik1 and -2 log lik2 would lead to the conclusion that the q variates in model that are additional to those in the other model improves the adequacy of the model. Under H_0, this test statistic is approximately distributed as χ^2 with *q* degree of freedom.

3.3.1.2 Akaike Information criterion (AIC)

Comparison between a number of possible unnested models *can* also be made on the basis of a statistic

$$AIC = -2\log \text{lik} + \alpha p$$

For each model, *AIC is* computed where p = df = number of unknown β-parameters in the model. α (2) is the predetermined constant used when n > 200 (α = 2 to 6). The better the model, the smaller the value of the *AIC* statistic. A 25% level of significance is used as a screening criterion for variable selection in this research work. (Bendel and Afifi 1977, Mickey and Greenland 1989 and Hosmer and Lemeshow 2000).

3.3.2 Model Selection Procedure

In order to model censored survival time as a function of a set of covariates, selection criteria are used to know which covariates to use. The following approaches were used to select the best model.

Collect's Model Selection approach

Collect recommended the approach of first doing a univariate analysis to "screen" out potentially significant variables for consideration in the multivariate model in order to identify the importance of each predictor.

Approach:

- Fit a univariate model for each covariate, and identify the predictors significant at some level $p1$, say 0.*2*0 (Hosmer and Lemeshow recommended $p1 = 0.25$).

- Fit a multivariate model with all significant univariate predictors, and use *backward* selection to eliminate non-significant variables at some level $p2$, say 0.10.

- Starting with final step (2) model, consider each of the non- significant variables from step (1) using *forward* selection, with significance level $p3$, say 0.10. Significant variables from *Step 2*, are forced into a multivariate model, using the INCLUDE option in SAS.

- Do final pruning of main-effects model (omit variables that are non-significant, add any that are significant), using *stepwise* regression with significance level $p4$. At this stage, you may also consider adding interactions between any of the main effects currently in the model, under the hierarchical principle.

Hierarchic principle means that if a model contains interaction term, the corresponding lower-order terms should also be included. Hierarchic principle is applied in all the model selection used in this research work.

Collett recommends using a likelihood ratio test for all variable inclusion/exclusion decisions. These approaches to covariate selection have been chosen since the use of one or more of them will yield, in the vast majority of model building applications, a subset of statistically and clinically significant covariates and applied all the automatic routines at once so this motivated me to interpret Collett in this research work.

Also, automatic routine of variable selection were also considered using the forward selection, backward elimination and stepwise procedure to compare if we arrive at the same covariates.

3.3.3 Model Diagnostic

After a model has been fitted to an observed set of survival data, the adequacy of the fitted model needs to be assessed. PH model checking procedures involve checking the assumption of proportional hazard and the use of residuals.

Graphical approach

There are various types of residuals to use in checking the adequacy of the fitted model. In this research work, efforts were made to diagnose using generalized (Cox-Snell), martingale, deviance, scale (Weighted) Schoenfeld and Schoenfeld Residuals and the Kaplan-Meier Plots.

- **Residual plots**

In generalized (Cox-Snell) residuals, estimated cumulative hazard for each individual at the time of their default or censoring should be like a censored sample from a unit exponential if the model is correct and the plot of log [-log S_e (t)] versus log (t) should yield a straight line through the origin with slope=1. Deviance residuals are calculated from Martingale residuals. They provide a solution to the asymmetric problem encountered with martingale residuals. For each person, the deviance residuals are defined as a function of the martingale residual introduced by Therneau, Grambsch and Fleming (1990), are much more symmetrically distributed about zero but not necessarily sum to zero after fitting model. The deviance residuals can then be plotted against the predicted log (HR) or each of the individual covariates. Allison (1995) states that deviance residuals are very informative for Cox models estimated by partial likelihood. Plot of deviance residuals are now used in a fuller assessment of the adequacy of the fitted Cox regression model. Scale (Weighted) Schoenfeld Residuals are more useful than the un-weighted version (Schoenfeld Residuals) because they are more like the typical OLS residuals. Thus, if the model fits well, they are expected to be random and symmetric around zero. The weighted residuals can be used in the same way as the unweighted ones to assess time trends and lack of proportionality.

- **Log-cumulative hazard plot**

A log-cumulative hazard plot, that is, a plot of the negative logarithm of the estimated survivor function against the logarithm of the survival time, will yield parallel curves if the hazards are proportional across the different groups. Also, the same condition works for the Kaplan-Meier plot of each covariate and the plots of weighted Schoenfeld residuals against time and Kleinbaum (1996) suggests that PH is assumed when weighted Schoenfield residuals clearly increases and decreases over time and that OLS regression line could be fitted to see if the slope is significant.

Formal Tests

- **Test based on time-dependent covariate**

This formal test is used to detect any time dependency in particular covariates, after allowing for the effects of explanatory variables that are known. Time-dependent covariates are covariates whose values change over time generated by creating interactions of the predictors and a function of survival time (time, or log (t)) and this can be added, to examine the assumption of PH in Cox regression model. Kleinbaum (1996) suggests that if any of the time dependent covariates are significant then those predictors are not proportional and this indicates a violation of the proportionality assumption for that specific predictor. Consider a PH model with two covariates Z_1 and Z_2. The standard PH model assumes

$$\lambda\,(t;\,Z) \;=\; \lambda_0\,(t)\,\exp\,(\,\beta_1 Z_1 + \;\beta_2 Z_2\,)$$

However, if the log-hazards are not really parallel between the groups defined by Z_2, then you can try adding an interaction with time:

$$\lambda\,(t;Z) \;=\; \lambda_0\,(t)\,\exp(\,\beta_1 Z_1 + \;\beta_2 Z_2 + \beta_3 Z_2 * t)$$

A test of the coefficient β_3 would be a test of the proportional hazards assumption for Z_2. If β_3 is positive, then the hazard ratio would be increasing over time; if negative, then decreasing over time.

- **Test based on goodness of fit**

Fitted PH model is also tested using Scaled Schoenfield Residuals (Grambsch and Therneau, 1994). If the PH assumption fails with respect to covariate Z_j, then the two subjects i and i', would differ in the j-th covariate but have all other covariates equal, instead of

$$\log\frac{\lambda_i(t)}{\lambda_{i'}(t)} = (Z_{ij} - Z_{i'j})\beta_j\,,$$

that is log (HR) constant in time; $\log\dfrac{\lambda_i(t)}{\lambda_{i'}(t)} = (Z_{ij} - Z_{i'j})\beta_j(t)$ where $\beta_j(t) = \beta_j + \gamma_j(t)$

Here, the weighted Schoenfeld residuals are used to estimate γ_j (t) (*Grambsch and Therneau, 1994*).This function can also be explored visually by the Schoenfeld Residuals. This is a global goodness of fit test of whether $\gamma_j(t) = 0, i.e \beta_j(t) = \beta_j$, is performed for all the variables.

3.3.4 Remedial Measures

The assumption says that if the hazards are not proportional, this means that the linear component of the model varies with time. If one of the predictors was not proportional there are various solutions to consider.

3.3.4.1 Accounting for Time-varying Covariates

Having checked if the covariates changed over time using the time-dependent covariates test described in Section 3.3.3. If the assumption of PH is not satisfied, time-dependent variable is used to analyze the time-independent predictors not satisfying the PH assumption.

3.3.4.2 Stratified Analyses

If the proportional hazard fails on an overall basis but that they are proportional in different subgroups of the data, a stratified proportional hazards model is built. Suppose proportionality assumption holds on Z_1 and *proportionality* simply does not hold between various levels of a second variable Z_2.If Z_2 is discrete (with *a* level) with enough data, then stratified model is fitted:

$$\lambda\,(t;\,Z_1\,,\,Z_2) = \;\lambda z_2\,(t)\,\exp(\,\beta Z_1\,)$$

A stratified model can be useful both for primary analysis and for checking the PH assumption.

3.3.5 Missing at Random (MAR)

Rubin (1976) proposed a formal definition of the missingness mechanism, defined on the subject level, as the conditional distribution of the missingness indicator vector (r_i), given the outcome vector $\boldsymbol{y}_i = (\boldsymbol{y}_i^{(o)}, \boldsymbol{y}_i^{(m)})$, the covariates ($\boldsymbol{x}_i$, including treatment assignment), and a vector of specific parameters ($\boldsymbol{\psi}$), i.e.,

Missingness mechanism = $P(r_i \mid y_i, x_i, \boldsymbol{\psi})$

We can classify missingness mechanisms in longitudinal studies, by extending the taxonomy used by Little and Rubin (1987) for intermittent missing data and that used by Little (1995) for dropouts. For intermittent missing data, missingness mechanisms can be classified into one of four types:
(1) missing completely at random (MCAR), where the probability that responses are intermittently missing is completely independent of all covariates ($\boldsymbol{x}_i$), all observed responses ($\boldsymbol{y}_i^{(o)}$), and all missing responses ($\boldsymbol{y}_i^{(m)}$);
(2) covariate-dependent missing at random (CMAR), where the intermittent missingness depends only on covariates ($\boldsymbol{x}_i$), a situation that can be crucial to hypothesis testing when the covariate-dependent missing involves dependence on treatment assignment;
(3) outcome-dependent missing at random (OMAR), where the intermittent missingness does not depend on the missing data ($\boldsymbol{y}_i^{(m)}$), but does depend on the observed responses ($\boldsymbol{y}_i^{(o)}$), and may also possibly depend on covariates ($\boldsymbol{x}_i$); and
(4) missing not at random (MNAR), where the intermittent missingness depends on the unobserved responses ($\boldsymbol{y}_i^{(m)}$), and may also possibly depend on covariates ($\boldsymbol{x}_i$) and observed responses ($\boldsymbol{y}_i^{(o)}$).
For dropouts, the mechanisms have similar classifications, though we assume that dropouts depend only on covariates, on previously observed responses, and on the first response where dropout begins.

Multiple Imputation Method

Rubin (1978) formally introduced Multiple Imputation (MI) and is used by imputing more than one value for each missing observation drawing a random sample of the missing values from its distribution and uncertainty due to imputation is introduced into the analysis. MI requires the missingness mechanism to be MAR and by re-combining estimates of parameters and covariance matrices will results in efficient and unbiased estimates and correct inference. It shares with single imputation the ability to used complete case analysis. Having known that assumption on the MCAR may not be realistic, under MAR assumption, analyses based on the direct likelihood are valid, then adoption of MAR are rarely justified. Here, many analyses could be used but focus is on multiple imputation since is used when there is a combination of missing covariates and missing outcomes. (Molenberghs and Verbeke 2005). Multiple imputation inference involves three distinct phases:

- The missing data are filled in M times to generate M complete data sets.
- The M complete data sets are analyzed using standard statistical analyses.
- The results from the M complete data sets are combined into a single inferential result.

Using the imputed data, inference about the parameter β is made by $(\beta - \hat{\beta}) \sim N(0,U)$ where U is the within imputation variance ($\hat{\text{var}}(\hat{\beta})$) .The M within-imputation estimates for β are pooled to give the multiple imputation estimate as

$$\hat{\beta}^{*} = \frac{\sum_{m=1}^{M} \hat{\beta}^{m}}{M}$$

Normal-based inferences for β could be based upon $\left(\beta - \hat{\beta}^{*}\right) \sim N(0,V)$ where

$$V = W + \left(\frac{M+1}{M}\right)B,$$

$$\text{W} = \frac{\sum_{m=1}^{M} U^{m}}{M},$$

is the average within imputation variance, and

$$\mathbf{B} = \frac{\sum_{m=1}^{M}\left(\hat{\beta}^{m} - \hat{\beta}^{*}\right)\left(\hat{\beta}^{m} - \hat{\beta}^{*}\right)'}{M-1}$$

is the between imputation variance. If the inferences are combined into a single one, then the average within imputation and between imputation variances are used to generate the confidence intervals (Rubin, 1987).

4. RESULTS

The data has been described and methodology in Section 2.0 and 3.0. Section 4.1 explores the data showing how the patients defaulted over time, the effects of the outcome response on the continuous and categorical covariates while Section 4.2 analyzes the original set of data, having organized the data horizontally using complete case analysis, and the missingness process using Multiple Imputation method is analyzed in Section 4.3 and loss to follow-up is analyzed in Section 4.4 after correcting for missingness.

4.1 Exploratory Data Analysis

The data comprised of 1167 observations obtained from five different datasets with some observations missing irrespective of the datasets. The outcome responses are defaulters and loss to follow-up. In ART-delivery programme in ITM, Antwerp, Table 4.1 shows that 13.97% of 1167 patients defaulted while 8.14% were lost to follow-up from the cohort study. The percentage of patients that defaulted is more than the patients that were lost to follow up. Most of the patients were censored in both outcomes.

Table 4.1: Summary of the Number of Event and Censored Values

Outcome Response	Event Indicator	Total	Percent
Defaulters	Censored	1004	86.03
	Event	163	13.97
Loss to follow-up	Censored	1072	91.86
	Event	95	8.14
Total		1167	100%

As shown from Table 4.2, defaulters and loss to follow-up have the same event for some period of years except in 2006 when the data was censored. Well, this is expected since a patient that is lost to follow-up has also defaulted on the condition that they do not come back at least one year before the last medical contact (31 December 2006) and higher number occurred in 2006 because of the patients that are away for more than one year at anytime during the treatment period and such patient may or may not be lost to follow-up. Also, out of 163 defaulters, more than half of them defaulted in 2006 alone (53.99%) while 3.68% defaulted in the first year of ART.

Table 4.2: Table of Defaulters and Loss to follow-up over a period of year

	YEAR						
EVENTS	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**Total**
Defaulters (%)	6(3.68%)	7(4.29%)	14(8.59%)	22(13.50%)	26(15.95%)	88(53.99%)	163(100%)
Loss to Follow-up (%)	6(6.32%)	7(7.37%)	14(14.74%)	22(23.16%)	26(23.37%)	20(21.05%)	95(100%)

Figure 4.1 presents the defaulters rate during a period of years at the HIV out-patients clinic and how it evolves over time. There is an increasing trend over the years in form of S-Shaped curve that begins with a slow start then followed by steep growth from 2005.

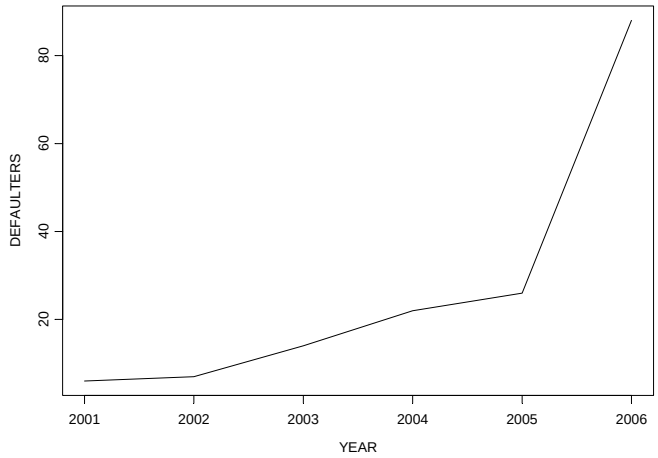


Figure 4.1: Plot of the Defaulters rate evolving over a period of time

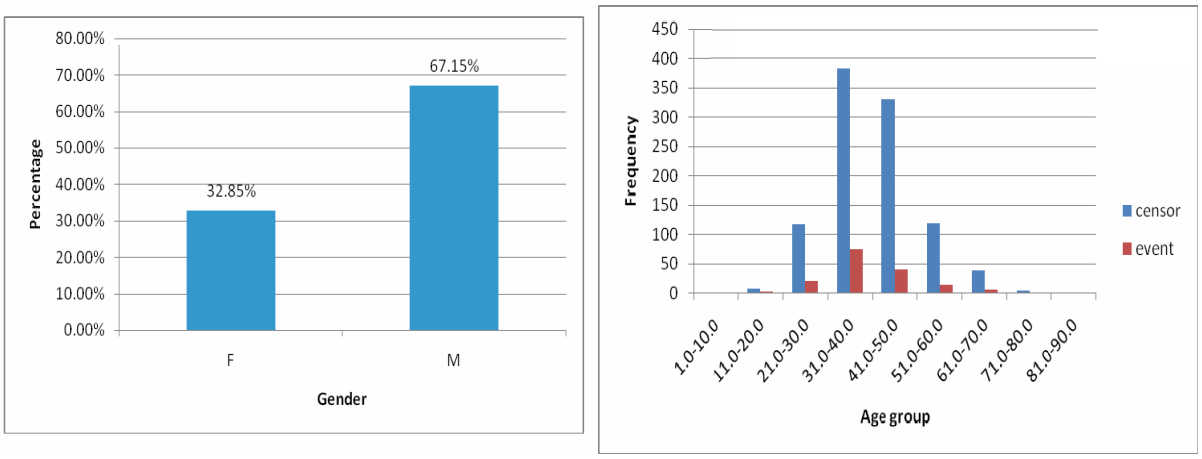
The summary statistics and the graphical illustrations of the categorical and continuous covariates are as follows:

Table 4.3: Descriptive Statistics of Continuous Variables by Gender (age, Average CD4 and Average Viral Load)

Variable	Sex	N	Maximum	Minimum	Mean	Std.Dev.	Percentage	Total
Age	F	383	80	17	37.790	9.73	32.82	1166
	M	783	81	18	42.560	9.76	67.10	
Average CD4	F	383	1500	-4000	1.936	273.55	32.82	1166
	M	783	3000	-6000	8.896	313.71	67.10	
Average Viral Load	F	383	1.2	-0.10	0.002	0.07	32.82	1166
	M	783	0.6	-1.50	-0.007	0.08	67.10	

Table 4.3 reveals that the males were over-represented in the groups with 783 males and 383 females. The minimum and maximum age for male patients were respectively 18 and 81, the youngest and oldest patients. The female median age is 37years and that of male is 42years.On the other hand, the minimum and maximum ages of females were respectively 17 and 80 with mean age of 37.79 years and standard deviation of 9.73. The average CD4 for male is 8.896, with a maximum of 3000 while female's average Viral Load is 0.0021.

(a) Histogram of Gender **(b) Multiple Bar-chart of age group by defaulters**



(c) Multiple Bar-Chart of Gender by defaulter **(d) Multiple Bar-Chart of Gender by Loss to follow**

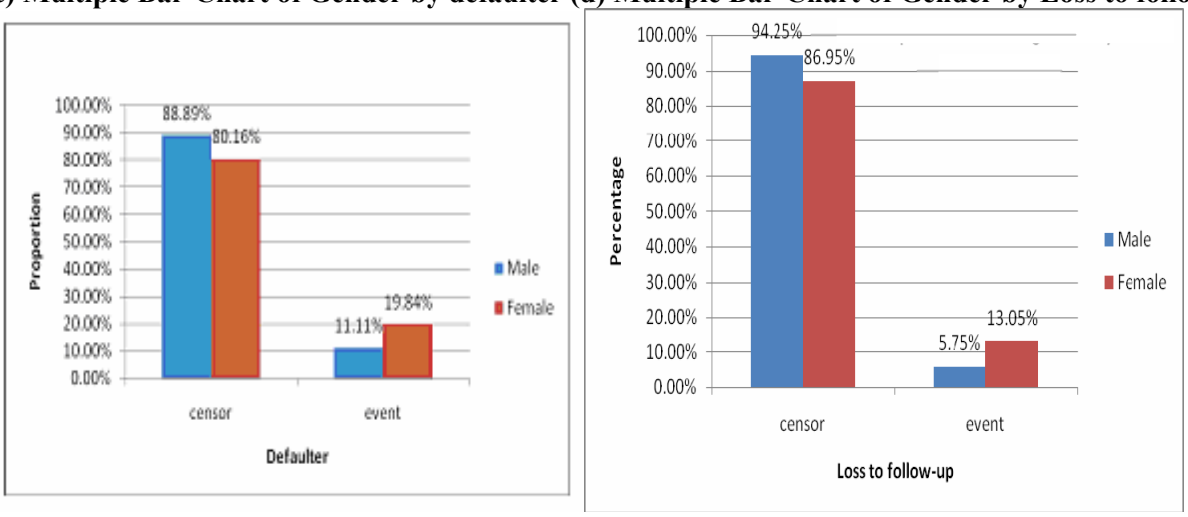


Figure 4.2:Plots showing the effect of Gender and Age group on the outcome responses.

Figure 4.2a reveals that the female patients have the lower percentage than the male patients in the study and we could see that patients below 50 defaulted more than patients in the other age categories as shown in Figure 4.2b. The percentage of female defaulters in Figure 4.2c is more than the male while most of the male patients were censored but in Figure 4.2d reveals that 13.05% of female were lost to follow up among the female patients, more than the male patients.

The categorical variables were explored to see the patients that defaulted, multiple-bar charts and Kaplan-Meier plots were displayed. Log-rank test of equality across strata (non-parametric test) was used to explore whether or not to include the predictor in the final model.

Table 4.15 in the appendix shows that the percentage of female that defaulted is more than male but averagely, female defaulted more than male. We could see that the patients with heterosexual contacts witnessed higher number of events while patients from the occupational risk and blood transfusion have the least percentage of defaulters, occupational risk defaulted most. 36.16% of patients were missing in the risk group. In all, 13.70% of patients defaulted in clinical stage and

most of the patients in Aids clinical stage were censored and the patients with Symptomatic clinical stage defaulted most. Hetero sex preference has the highest patient default and also in percentage. Meanwhile the only patient with unknown sex preference was censored. The proportion of patients that defaulted differs among continents, ranging from 12% in Europe to 24% in America. Patients from Europe had the highest percentage among all the patients (51.07%) but having the least percentage of defaulters while patients from Oceania had the least percentage (1.47%) among all the patients but they defaulted most.12.87% of patients defaulted from the entire origin group. The table shows that the oldest male patient followed up throughout the study period. This indicates that patients within the age group (31-40) defaulted most in both sex having the highest number of patients throughout the study, age group (61-70) have the least female defaulters.

From Table 4.15 in the appendix, we found no much difference between the defaulted patients with high and low viral load. Of the 1167 patients included in this analysis, approximately 7.52% (69) defaulters had at least one viral load below 400 copies/ml during the period of the study. Note that patient with high viral load might default as a result of discouragement if there is no improvement over time having known that high viral load at first consultation means the disease is really active.

Patients with CD4 count $\geq$ 0.2 defaulted most and as a result, this might not be an important covariate since this table indicated that higher percentage of defaulted patients are not seriously sick based on the CD4 count. Also, all the patients that were not on ART defaulted from the study. Half of the patients that were on ART defaulted as well.

Table 4.4: Summary of the Number of Event and Censored Values for all continuous variables

Continuous Covariates	**Total**	**Event**	**Censored**	**Percent Censored**	**Percent Defaulted**	**Percent Missing**
Age	1164	162	1002	86.08	13.92	3 (0.26%)
Average CD4 count	1151	161	990	86.01	13.99	16 (1.37%)
Average Viral Load	935	143	792	84.71	15.29	232 (19.88%)

Table 4.4 presents the summary of the number of event and censored values of the continuous variables. Only 13.92% of the patients out of the 1164 patients defaulted. 13.99% of patients with average CD4 count defaulted. Due to missingness in the average viral load of patients, 935 observations were used and 15.29% defaulted.

The Log rank statistic derived here is the CMH type log rank test testing for conditional independence H0: OR = 1 since we used "STRATA" statement in SAS. The log rank tests for homogeneity indicate a significant difference between the gender (p=0.0002) for the log-rank test, where female patients defaulted significantly more than those of male patients. This has shown that there is conditional dependence since the OR is not equal to one, the defaulters are less likely in male patients. The result is expected as it was illustrated from Figure 4.3a. Having met the p-value of 0.2 - 0.25 or less (Hosmer and Lemeshow 2000*)* as shown in the inset of each KM plot in Figure 4.3 and Figure 4.6 in the appendix, the log-rank test of equality across strata p values indicate that there is an evidence of a significant difference among the survival curves for all these groups at 25% level, thus gender, risk group, clinical stage, sex preference,

origin group, ART, viral load and age group might be included as potential candidates for the final model and there could be association between these covariates and defaulters. The same sets of covariates were significant at 25% level of significance as displayed in Table 4.5 with the univariate analysis of the PH model.

Table 4.5: Univariate Cox PH model for all Categorical variables

Variable	DF	Parameter Estimate	Std. Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Defaulters						
Sex	1	0.5843	0.1570	13.8411	0.0002	1.7940
Sexpreference	1	0.2446	0.1586	2.3805	0.1229	1.2770
Clinical stage	1	-0.5071	0.1625	9.7435	0.0018	0.6020
Risk group	1	0.1488	0.1139	1.7059	0.1915	1.1600
ART	1	-3.8849	0.1766	484.0119	<.0001	0.0210
Viral Load	1	1.9053	0.1589	143.6533	<.0001	6.7210
Origin group	1	0.2716	0.0814	11.1261	0.0009	1.3120
Loss to follow up						
Sex	1	0.8244	0.2055	16.0957	<.0001	2.2810
Sexpreference	1	0.5227	0.2163	5.8403	0.0157	1.6870
Clinical stage	1	-0.6721	0.2055	10.6932	0.0011	0.5110
Risk group	1	0.1003	0.1602	0.3916	0.5315	1.1050
Viral Load	1	2.6821	0.2398	125.1398	<.0001	14.6160
Origin group	1	0.3118	0.1021	9.3258	0.0023	1.3660

For defaulted and lost to follow-up patients, the chi-squared test for age has a p-value <. 0001 so we reject the null hypothesis and conclude that age is significantly related to survival time. Average CD4 count might not contribute anything to the model if added since it is not significant but could be included being important variable. The Chi-squared test for average viral load has a p-value of 0.0217 this is significant enough to reject the null hypothesis of not including it in the model and this makes these covariates to be a potential candidate in the final model. For ART, no patient was lost to follow-up so there was no output.

Table 4.6: Univariate Cox PH model for all Continuous variables

Variable	DF	Parameter Estimate	Std. Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Defaulters						
Age	1	-0.0471	0.0095	24.8713	<. 0001	0.954
Average CD4	1	-0.0001	0.0002	0.2277	0.6332	1.000
Average Viral Load	1	1.3269	0.5782	5.2666	0.0217	3.769
Loss to follow up						
Age	1	-0.0593	0.0124	23.0548	<. 0001	0.942
Average CD4	1	0.0004	0.0003	2.0164	0.1556	1.000
Average Viral Load	1	0.5519	1.1139	0.2454	0.6203	1.736

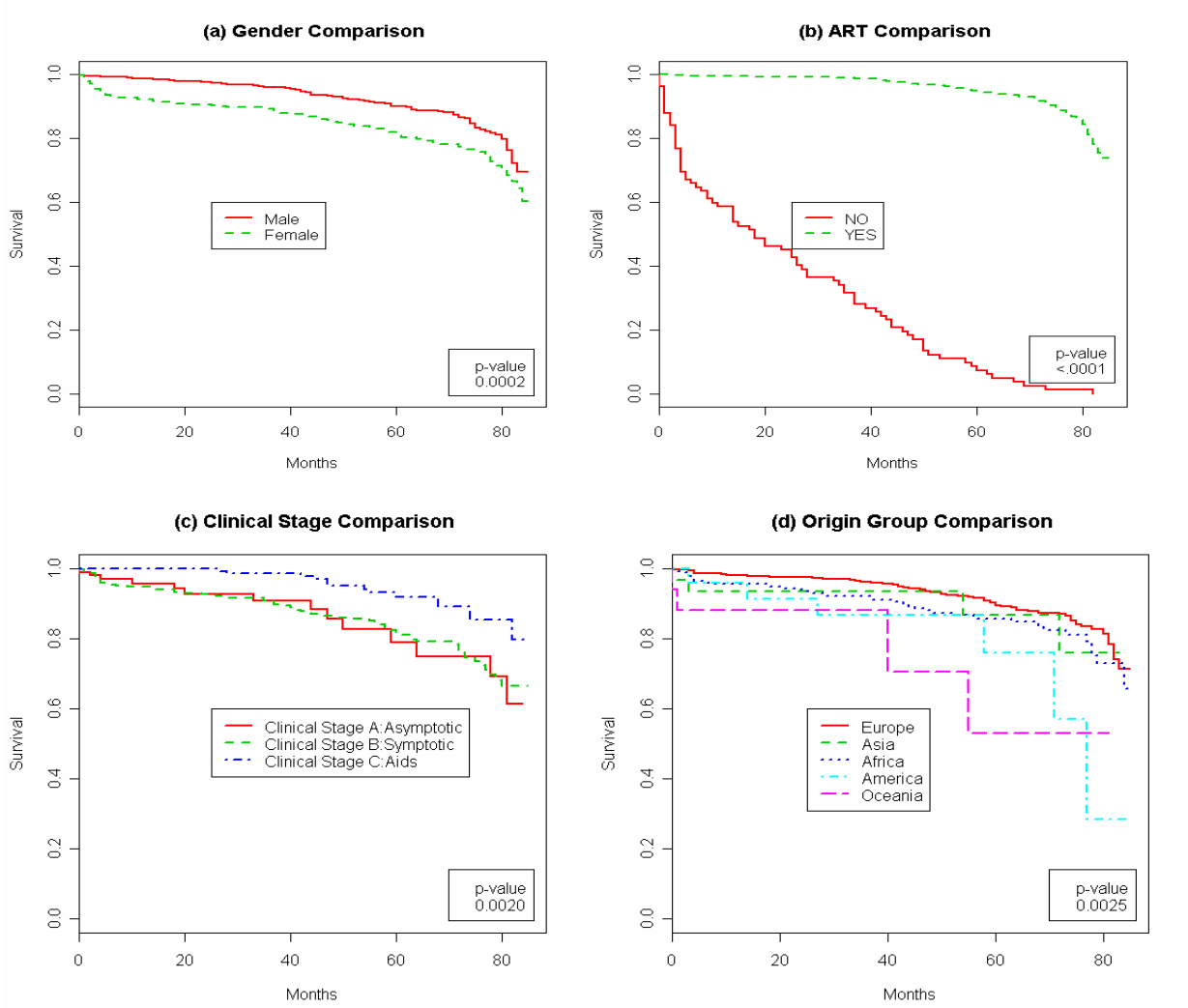


Figure 4.3: Graph of estimated survivorship function of Gender, ART, Clinical Stage and Origin group comparison.

Figure 4.3a shows that the survival function for each group of gender is not parallel but they are separate except at the very beginning of the study time. The estimated survival curves did not cross each other, evidence that the hazards are proportional. In general, the log-rank test places more emphasis on the differences in the curves at larger time values. This is why we get such a small p-value even though the two survival curves appear to be very close together. The difference in displayed survival curves reinforces the conclusions that the patients in the male group defaulted more than the patients in the female group. In Figure 4.3b, hazard proportional assumption might be justified since the two estimated survival curves do not cross each other at any point, the tests for equality over strata confirms it with highly significant value but conclusion cannot be reached yet until the formal test is done.

It is important to know that patients on ART at last consultation means that the patients were on treatment before defaulting while those not on ART, were on consultation at that time of default i.e not on treatment at that time. The criteria used here is that if the date of last consultation is more than the date of last treatment, then the patient is not on ART at last consultation. It is necessary to know how many patients were on treatment when they came for the last

consultation because once the patients defaulted, they might soon fall short of drugs and may be sick again. Patients not on ART at last consultation decreases over time throughout the study while patients on ART were constant for some months before it finally defaulted showing that those patients not on ART tended to default sooner. This Figure shows that the estimated survivor function for patients on ART is always greater than that of the patients without ART. The graph in Figure 4.3c shows that the three groups are not parallel. The asymptomatic and symptomatic clinical stages overlap for most of the graph but the shape of the Aids clinical stage curve is quite different, this lack of parallelism could pose a problem when we include this predictor in the Cox proportional hazard model since one of the assumptions is proportionality of all the predictors. The survival rates of the asymptomatic clinical stage patients and the symptomatic clinical stage patients decrease gradually. From Figure 4.3d, we could see that the patient from America defaults rapidly and this caused the percent surviving decreases to almost 25%. When the first patient defaulted the percent survival was constant for about 40 month. When the next patient defaulted, the percent survival dropped again to 70%. In all, the origin group curves are almost close to one another except Oceania. Assumption of hazard proportionality does not hold here since the survivorship curves cross.

Figure 4.6c in the appendix displays the risk group curves which are much closer to one another. The survival rate of the occupational risk patients' decrease to almost 80% in 40 months. Shapes of the blood transfusion curve and the heterosexual contracts curve are quite different, although both decrease in the same pattern. Having seen that the estimated survival curves cross one another, this is evidence that the hazards are not proportional in the patients with risk group.
From Figure 4.6d in the appendix, we see that the survival function for each group of sex preference are not perfectly parallel but that they are separate except at the beginning and end of the study time for Hetero and Homo sex preference. The straight line that appears for unknown is evidence that the survivor was 100% since no patient defaulted at that stage. Continuous downward steps in the homo and hetero sex preference curves are caused as a result of the higher number of defaulters that made the curves difficult to interpret.

Also, we could see that the survival curves of the two viral load groups are not really parallel both overlap as shown from Figure 4.6a in the appendix and that there are two periods ([0, 20] and [60, 70]) where the curves are very close together. This would explain the high p-value from the log-rank test. Also, this lack of parallelism could pose a problem when we include this predictor in the Cox proportional hazard model since one of the assumptions is proportionality of all the predictors.

Prior knowledge shows that since the lower the CD4 count the more progressive the HIV disease so patient with CD4 Count < 0.2 could reflect a risk of opportunistic infections [10]. The defaulted patients with lowest CD4 count were checked to see whether patients who were very sick are more prone to default. Figure 4.6b in the appendix shows that the estimated survivor function for patients below CD4 count is more than that of the patients above CD4 count. In particular, patients with above CD4 count appear to default more than the patients with below CD4 count. In general, the log-rank test places more emphasis on the differences in the curves at larger time values. This is why we get such a small p-value even though the two survival curves appear to be very close together for time close to 80 months.

Since age might be related to survival from past literature, in order to do a preliminary analyses that can easily understood, age was divided into several groups of interest. In Figure 4.6e in the appendix, information on the estimated survival function of age group shows that all the age group curves are much closer to one another. Since the estimated survival curves cross one another, this is evidence that the hazards might not proportional in the patients with age group. Figure 4.7 in the appendix shows that higher number of the male patients was censored and had event. This is due to the higher percentage (67.15%) male had among all the patients.

From Table 4.16 in the appendix, information on the doctors whose patients defaulted is given. In all, 17 doctors attended to all the 1167 Patients. The highest number of patients treated by a doctor is 193 from 2004-2006 and illustration from the dataset shows that patients went for consultation from 2003 –2006. Doctor code 1775 had 156 patients in which 32 patients out of them defaulted (20.51%) having highest percentage of defaulters. 62.64% of the 1167 patients came for consultation in 2003 and 10 out of the 17 doctors attended to them.

4.2 Statistical Analysis

Model Selection procedure has been introduced in Section 3.3.2. Section 4.2.1 analyzes the original set of data (i.e. without artificially removed subjects) and the model selection diagnostic was checked in Section 4.2.2 using the PH assumption and Stratified analysis previously described in Sections 3.2.5 and 3.2.6.

4.2.1 Model Selection Procedure

Firstly, model including all the predictors with a p-value of less than 0.2 - 0.25 in the univariate analyses are considered i.e every predictor will be in our model. In the model statement, variable that contains the information about time **(newmonthts)** was specified and the variable containing information about censoring (**defaulters)**. In this model we therefore specify zero since the coding for **censor** is that **defaulters** = 0 indicates that the subject has been censored and **defaulters** = 1 indicates that the subject experienced an event. This is used in fitting the Cox's proportional hazard model.

Collett's approach for model selection is in four major steps, treating all variables equally. We use an entry probability for the univariate predictors of α = 0.25, and a significance level of 0.10 for all other entry/exit probabilities. The Likelihood Ratio Test was used for all variable inclusion/exclusion decisions and AIC values to compare models.

Univariate analysis was fitted on all variables as in Tables 4.5 and 4.6 and the significant variables at 25% were the potential variables for the next stage of the selection procedure. These variables were used in the next stage in a Multivariate model using *backward* selection procedure with exit probability 0.10. Average CD4 (p valve = 0.6332) was removed from the multivariate model at this stage. The non-significant variable (Average CD4) in the first stage was then added to the third stage with the significant variables from the second stage and refitted using for*ward* selection procedure to select the final model at 10% level for entering and exist from the model. The variables gender, sex preference and ART, from the final model of second s*tage*, are forced

into a multivariate model of the third stage using the **INCLUDE** option in **SAS.** The variable that is significant is retained in the model and the one that is not significant is removed before adding another variable to the model. Finally, pair wise interactions among the final variables in stage 3 were considered and multivariate model fitted using Stepwise selection Procedure. All the non-significant variables from stage 3 were omitted here. Hierarchical principle was followed in this procedure by including main effects with their interactions in the model, even when not significant. The interaction with highest non-significant p-values were removed one at a time and the model fitted until all the variables in the model could no longer be removed at 5% level of significant. Note that the AIC of the final fitted model is lower than the AICs of the entire univariate model indicating that this model is 'better' than the univariate models.

Table 4.7: Final Collett's model before the intermediate model application

Variables	Parameter Estimate	Standard Error	P-value	Hazard ratio
Sex	0.6212	0.1601	0.0001	1.861
ART	-3.9592	0.1828	<. 0001	0.019
SexAvgcd4	-0.0004	0.0002	0.0090	1.000

Three variables were significant having considered all the covariates from stage 3 and their interactions but we need to check for the intermediate models. The likelihood ratio test was used to know which variables should be in the model. This test statistic has a chi-square distribution under the null hypothesis of no interaction effect, models 2 and 3 from Table 4.8 are compared. $\chi^2_{LR} = -2\log lik_1 + 2\log lik_2$ = 1444.380 - 1442.697 = 1.683 this gives a non-significant p-value of 0.1945. Therefore, the interaction can be dropped.

Table 4.8: Intermediate models from the stepwise procedure to the Collett's final model

Model Number	Variables	$-2Log(\hat{L})$	q	AIC
1	Sex ART sexavgcd4	1442.645	3	1448.645
2	Sex ART avgcd4 sexavgcd4	1442.697	4	1450.697
3	Sex ART avgcd4	1444.380	3	1450.380

Having considered the hierarchical principle, the result of the final model from the Collett's procedure is shown in Table 4.9. The estimate hazard for default obtained in the form of a Cox proportional hazards models is:

$$\lambda(t,Z) = \lambda_0 \exp(0.62104 gender - 3.95644 ART - 0.000259 AvgCD4)$$

Table 4.9: Final model from the Collett's model selection procedures

Variable	Estimate	Standard Error	P-value	Hazard Ratio
Sex	0.6210	0.1601	0.0001	1.861
ART	-3.9564	0.1825	<. 0001	0.019
Average CD4count	-0.0003	0.0001	0.0356	1.000

According to the HR for gender, male patients are 1.9 times at high risk as compared to female patients, while controlling other factors. ART estimated hazard ratio of 0.019 means that the death risk in the higher ART category is about 0.02 times the death risk in the category below. The 95% confidence interval for the sex effect (HR), which surrounds the point estimate of 1.861, though as expected because of the low p-value of 0.0001 shows that the confidence interval for HR does not contain the null value of 1.Also, sex in this model gives a more precise estimate of the HR unlike in other models which confirms the validity of this model.

The automatic selection criteria (forward, backward and stepwise selection criteria) were used to fit the model as the Collett's approach all giving the same final model but automatic selection criteria procedure has one drawback that they could only handle variables one at time.

Also, the model is selected in this way, covariates with a small number of missingness like age, gender, sex preference, ART and averageCD4 were first included into the model and covariates that were not significant among these five were removed and the next covariate with a lesser missingness was added one at a time to the remaining till a univariate analyses was done. The interactions of all the covariates that passed through the univariate analyses were modelled together. In conclusion, the same set of covariates we got from the model fitted using the Collett's approach and the automatic selection criteria procedures were also significant in this criterion but finally Collect's approach is adopted for use in this research work for the reason mentioned in Section 3.3.2.

4.2.2 Model Diagnostic

After arriving at the most parsimonious model, the next step in statistical analysis is to diagnose the model and see if it is appropriate enough to be used as a model to explain the overall survival of the patients.

Graphical Approach

One of the main assumptions of the Cox proportional hazard model is proportionality. Here, graphical approach is used to verify that a model satisfies the assumption of proportionality.

- **Residual Plots**

To know if the final PH model containing gender, ART and Average CD4 are well fitted, generalized (Cox-Shell) residuals was done by plotting log [-logS (t)] versus log (t) and this should yield a straight line through the origin with slope=1 if the model satisfies the PH assumption. From Figure 4.4, the residuals, especially for small values, do not yield exactly a straight line. Anyway, we do not trust too much about these results, since Allison (1995) claims that Cox-Snell residuals are considered as not very informative for the Cox model. The points seem to be on a line, which passes through the origin. Though, there is slight deviation from the straight line which is mostly noticed at the tail, it may be concluded that the model fits but until the formal test is done.

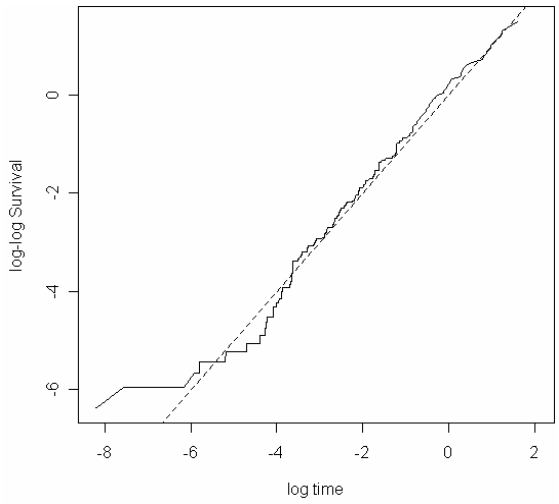


Figure 4.4: [-logS (t)] versus log (t)

The deviance residuals are calculated from Martingale residuals solving the asymmetric problem encountered with martingale residuals. The plot of the deviance residuals against the linear predictor and each of the covariates are presented in Figure 4.8 in the appendix. Thus, if the model fits the data well, they should be randomly scattered around zero. It looks as if the deviance residuals had not corrected the asymmetric problem of martingale residuals since the residuals appear not to be randomly scattered and symmetric around zero. Thus, it appears that the model seems not to fit the data well but formal test will explain more on it. Also, scaled (weighted) Schoenfeld residuals were used and are expected to be random and symmetric around zero if the model fits well data. The plots from Figure 4.9 in the appendix shows that the model is not well fitted for sex and ART since the Schoenfeld residuals were not random and symmetric around zero but were only random and symmetric around zero for average CD4.

- **Log-cumulative hazard plot**

To assess if the hazards are actually proportional to each other over time, the plot of log Λ = log [- log (^*S*)] vs. log (*t*) for two sub-groups is used. If the plot has parallel (lines or curves), then PH assumption holds. We compared log cumulative hazard, (log Λ (*t*)), between the covariate levels as displayed in Figure 4.5.

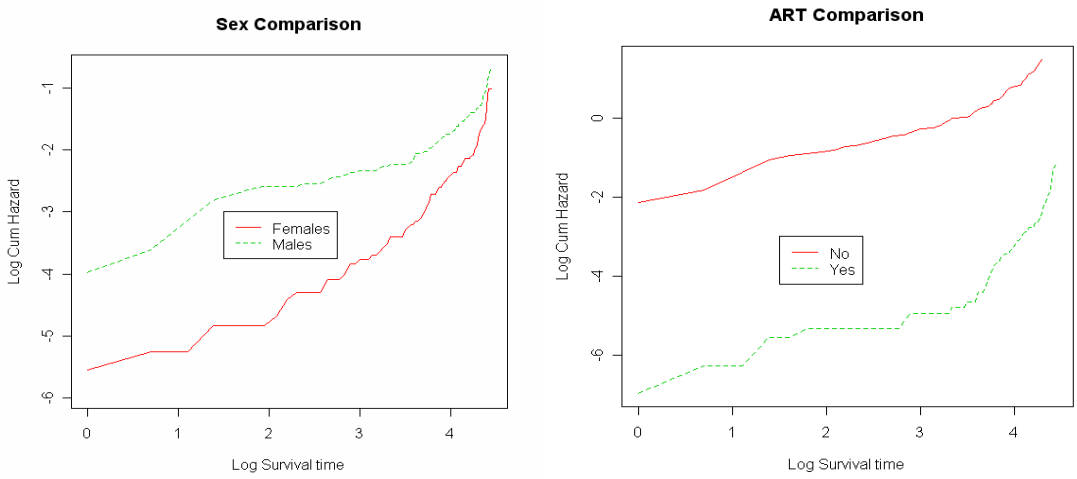


Figure 4.5: Assessment of the PH assumption by SEX and ART category

The plots seem to show that the hazard ratio between the two groups depends on time. Thus, it cannot be assumed that the proportional hazards assumption holds. However, this should be confirmed by formal tests.

Formal Tests

Some formal tests were done to check for the assumption of Proportional Hazard after the informal test through the graphical approach.

- **Test based on time-dependent covariate**

Proportionality is checked by testing the time-varying coefficient $\beta = \beta(t)$ including interaction terms between covariates *Zj* and log (*t*) (i.e. time-dependent covariates) in the model. Sex and ART were fitted with a term corresponding to an interaction between them. This interaction was modeled by including the time-dependent variable Sext and ARTt.

Table 4.10: Test statistics for proportional hazards based on interactions between covariates and time

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Sex	1	2.1955	0.5551	15.6453	<. 0001	8.984
ART	1	-5.2408	0.7349	50.8547	<. 0001	0.005
AverageCD4	1	-0.0007	0.0006	1.5643	0.2110	0.999
Sext	1	-0.4753	0.1550	9.3985	0.0022	0.622
ARTt	1	0.4008	0.2043	3.8478	0.0498	1.493
AverageCD4t	**1**	**0.0001**	**0.0001**	**0.7053**	**0.4010**	**1.000**

Table 4.10 shows the estimates from a Proportional Hazards (PH) model using the variables sex, ART and AverageCD4count and their interaction with time. Kleinbaum (1996) suggested that it can be assumed that the PH assumption holds unless there is strong evidence that the test for log (time)*covariate interaction term is significant. From Table 4.10, both Sext and ARTt are significant i.e do not satisfy PH assumption. Also, an overall chi-square test for the *covariate*time* interaction from the linear hypothesis result has a chi-square value of 13.2246 and a p-value of 0.0042 for 3 degrees of freedom. Hence, it can be assumed that the PH model does not hold since the overall effect of the interaction between *covariate* and *time* are significant. Our faith in this result is bolstered by the Kaplan-Meier curves we created during our univariate analyses. The curves for all the variables in the model were fairly separated, then cross which violates proportionality assumption.

- **Test based on goodness of fit**

A global goodness of fit test of whether $\gamma_j(t) = 0, i.e \beta_j(t) = \beta_j$, is also performed on the fitted Cox's PH model to check for the PH assumption. This function was explored visually as well for each fitted covariate as presented in Figure 4.10 in the appendix and the result from the formal test is shown in Table 4.11. This is done with the use of **cox.zph** in R language.

The graphical illustrations for this output is given in Figure 4.10 in the appendix with lowess smooth curve which do not deviate much from zero indicating that the assumption holds in the case of averageCD4. Kleinbaum (1996) suggests that it can be assumed that PH assumption holds unless there is strong evidence that the weighted Schoenfield residuals clearly increases and decreases over time. PH assumption does not hold for the sex and ART based on *Kleinbaum* suggestion since the residuals are concentrated around the slope. Also, the Schoenfeld residuals seem to show that the lowess smooth curve deviates much from zero indicating that the

assumption does not hold. Table 4.11 presents the results from the global goodness of fit test on all the covariates in the fitted model.

Table 4.11: Formal test for the Proportional Hazard assumption

Variable	Chi square	p-value
Sex	3.379	0.0660
ART	4.826	0.0280
AverageCD4	0.217	0.6411
Global test	7.635	0.0542

The results confirm that the proportional hazard assumption should be rejected at 5% level of significance for ART with p-value smaller than 5%. Though it appears that sex was not significant, therefore the assumption holds but a p value of 6% level of significance could still not be trusted. Only the average CD4count with a larger significant level of 64% could affirm the proportionality assumption. In conclusion, this global test seems to be good though with PH borderline.

4.2.3 Remedial Measures

This is necessary since some of the predictors do not satisfy the proportional hazard assumption.

4.2.3.1 Accounting for Time-varying covariate

Having confirmed that PH assumptions did not hold for the time-dependent covariate, here we accounted for the covariates as they vary with time.

4.2.3.2 Stratified Analyses

This was done since the PH assumption holds for a covariate and did not hold for some other covariates leading to stratified Cox PH model. Here, two models are involved, the Cox's PH model is compared with the second model where the covariates are stratified by gender and ART thereby making PH assumption for unstratified covariates.

Table 4.12: Result of the Cox's PH model and Stratified PH model for Sex

Model Type	Variable	coef	exp (coef)	Se (coef)	z	HR	p-value
Cox'sPH model	Sex	0.6231	1.8646	0.1599	3.89	1.8647	9.8e-05
	ART	-3.9486	0.0193	0.1814	-21.77	0.0193	0.0e+00
	AverageCD4	-0.0003	0.9997	0.0001	-2.10	0.9997	3.6e-02
Stratified Cox's	ART	-3.8841	0.0206	0.1817	-21.37	0.0206	0.0000
PH Model	AverageCD4	-0.0003	0.9998	0.0001	-2.02	0.9997	0.0430

Having checked PH assumption for each categorical covariate by repeating the stratification for each, comparing the results with the Cox's PH model and stratified PH model by gender is shown in Table 4.12.Since variables ART and AverageCD4 are included in the stratified Cox's PH model, we can estimate the effect of each variable adjusted for the other variable and the sex variable but not possible to obtain a HR value for the effect of sex adjusted for the other two variables since sex is not included in the model. Here, we control for sex by stratification and the results show that the estimates for the other covariates are mostly the same for the two models, the increase in the standard error is negligible for both stratification which implies that stratification might not be necessary. The plot in Figure 4.11a in the appendix shows that we

cannot assume that the proportional hazards assumption holds even with the stratification. The same conclusion was reached from the stratification by ART (Table 4.17 in the appendix) and Figure 4.11b in the appendix.

4.3 Missing at Random (Imputed Data)

Multiple imputation procedure is used to fit the model of the incomplete dataset, which made the analysis straightforward. After fitting the Cox's PH model with the imputed data, the assumption of Proportional Assumption is checked using formal tests.

The model is selected using Collett's method, which was previously introduced in Section 3.3.2. Having undergone all the procedures involve in Collett's method, Table 4.13 presents the final model after correcting for missingness showing all the significant covariates.The covariates (ART and Sex) present in the fitted model before and after correcting for missingness.

Table 4.13: Final Collett's model after correcting for missingness

Variables	Estimate	Std. Error	95% Confidence Limits		DF	Pr > |t|
Age	-0.0341	0.0087	-0.0512	-0.0170	34014	<. 0001
Average Viral Load	1.3698	0.7267	-0.0586	2.7983	423.11	0.0401
ART	-3.9345	0.1829	-4.2931	-3.5759	5.9E7	<. 0001
Sex	0.5195	0.1627	0.2005	0.8385	595147	0.0014

Generalized (Cox-Shell) Residuals was used to check the fitted model if it is satisfactory. Figure 4.12 in the appendix shows a plot of log [-logS (*t*)] versus log (*t*). The points are on a line, which passes through the origin. Though, there is slight deviation from the straight line, it may be concluded that the model fits but formal test will justify it better. Proportionality is checked by testing with time-dependent covariates in the model and also through the use of global goodness of fit. The final covariates were fitted with a term corresponding to an interaction between them, which was modeled by including the time-dependent variable. Table 4.18 in the appendix presents the estimates from a Proportional Hazards (PH) model using the variables age, sex, ART and Average Viral Load and their interaction with time. From Table 4.18 in the appendix, all the time-dependent time were not significant. Also, an overall chi-square test for the *treatment*time* interaction from the linear hypothesis result had a chi-square value of 0.3007 and a p-value of 0.8012 for 4 degrees of freedom. Hence, it can be assumed that the PH model holds since the overall effect of the interaction between *treatment* and *time* are not significant. Table 4.14 presents the results from the global goodness of fit test on all the covariates in the fitted model. The results confirm that the proportional hazard assumption cannot be rejected for any of these covariates since they are all not significant. In conclusion, the global goodness of fit test confirms the assumption of PH, seems to be good, with the p value greater than 5% (0.543).

Table 4.14: Formal test for the Proportional Hazard assumption

Variable	Chi square	p-value
Age	2.86e-01	0.853
Average Viral Load	1.97e-02	0.901
ART	7.22e-02	0.389
Sex	3.80e-04	0.722
Global test	2.29e+01	0.543

4.4 Loss to follow-up

As discussed in Section 2 that there are two outcome responses (Defaulters and Loss to follow-up) in this analysis and having modeled defaulters, loss to follow-up is also analyzed after missingness has been corrected but in order to save space, the analysis is summarized in this section.

Having followed the Collett's procedures and undergone intermediate hierarchical model selection procedure was used to justify the best model, imputation procedure was used to get the final Cox's PH model (Age, Sex and Clinical Stage) which is presented in Table 4.20 in the appendix. The three covariates were significant at 5% level of significance, sex and age were among the covariates in the model fitted when analyzed with both defaulters and loss to follow-up. The estimates and standard error of the two covariates are larger in loss to follow-up than defaulters. This might be due to the lower percentage of loss to follow-up (8.14%) in the study.

The model is diagnosed using generalized (Cox-Shell) residuals but there is slight deviation from the straight line, it may be concluded that the model fits and the formal test justified it better using time-dependent covariate and goodness of fit test. The two formal tests show that the PH assumption holds for all the covariates at 5% level of significance as presented in Tables 4.21 and 4.22 in the appendix meaning that the model is a good fit.

5. DISCUSSION

This research tries to look at the rate at which patients default and loss to follow-up over time, their risk factors in the HIV/STD unit of Institute of Tropical Medicine, Antwerp. The defaulter and lost to follow-up rates were estimated using survival analysis method and the risk factors for defaulters were assessed using the Cox proportional hazard models.

Five datasets (patients, treatment, consultation, viral load and CD4count) were merged together, having 1167 observations were used in this analysis. Many researchers use series of mathematical and statistical models to check for defaulters rate over time, possible causes of patient default and the risk factors associated with their default.

This study shows that after carrying out a univariate analysis on the continuous and categorical covariates, the defaulter's rate vary from one covariate to another. Majority of the patients were male but in proportion, female defaulted more than male with 19.84% while 13.05% of female were lost to follow-up. Young female patients defaulted more than adult patients having known that both male and female differ in the study. Heterosexual contacts patients defaulted most in the study which reflects past research papers concerning HIV while the asymptomatic clinical stage had many defaulters with 16.59%, this is expected since in this stage the immune system becomes severely damaged by HIV. The treatment failure at last consultation shows that 37.75% from 94 patients with high viral load, this might be dangerous since with this viral load their disease will be very active and HIV could be transmitted easily. Note that this treatment failure might occur because patients are not using their drugs as prescribed. Since patients with high CD4count defaulted most during the period of the study, a risk of opportunistic infections might not be reflected so most of the defaulted patients are not very sick since their Nadir CD4count is above 0.2. All the patients that were not on ART at last consultation defaulted from the study, even defaulted sooner than those on ART. The demerit of this is that they might soon fall short of drugs and may be sick again and less than 15% patients from average CD4count and average Viral Load defaulted from the study.

The Cox's PH model fitted using complete case analysis shows that sex, ART and average CD4count were significant but they do not satisfy the assumption of proportional hazard except the average CD4count.The residual plots and the test of time-dependent covariate indicated that the model were not well fitted since they are significant after considering the time-varying covariate term except for the average CD4count that satisfies the condition. The result from stratified analysis even shows that stratification might not be necessary since there is no much change in the parameter estimates and standard errors.

The model fitted after imputation satisfies the assumption of proportional hazard. The formal test applied to the model shows that the time-dependent covariates were not significant and the global fit test also shows that all the covariates were not significant which justifies the PH assumption. Though, the model fitted unlike the previous model shows that there is no need to test further with the use of stratified analysis since the assumption of PH was met and covariates differ in both models except Sex and ART that are present in both with different parameter estimates and standard error.

Exploratory results from loss to follow-up analysis on all the covariates show that patients that defaulted are nearly the same to those lost to follow-up, their difference is less than 6%. Female lost to follow-up is significantly difference from male since female had higher lost to follow up percentage than male. Age, sex and clinical stage were significant and the PH assumption held after correcting for missingness.

This research work has several limitations, most times, we do not know the causes of patients default from Antiretroviral Therapy may be death, loss to follow-up, followed in another hospital but many approaches and attempts had been made in the past by scientist and researchers to study defaulters rate in a cohort of HIV infected patients In their papers, the true outcome status of patients at the ART tracing visit of the clinic were given so this made it convenient to know the causes of default and most of the reasons for default were death, alive on ART at a different clinic and some stopped therapy as a result of high cost of transport to the clinic, religious beliefs, persuasion by relatives to stop ART and other reasons. Some could not be traced because of an incorrect address in the ART register. To address this issue is difficult since there is no availability of this information in my dataset so nothing could be done. Researchers noted some of these points as a way to decrease the loss to follow-up or get patients back to medical system. Issues like reducing the cost of transport to the clinic, continuous home visit are mentioned [11].

6. CONCLUSION

There is increasing trend in the rate at which the patients defaulted during the last five years through its evolution over time. It was observed that the model does not satisfy the proportional hazard assumption even with time dependent covariates but the global goodness of fit test shows that the model is fitted at a borderline significant level. Allison (1995) claims that too much emphasis is always put on testing the PH assumption, and not enough to other important aspects of the model. He said further that if proportional hazard doesn't exactly hold for a particular covariate having fitted the PH model, then we are getting a sort of an average hazard ratio, averaged over the event times and the fitted model is not such a bad estimate. The averageCD4 count when tested with time-dependent covariate, satisfy the proportional hazard assumption. Exploratory analysis also confirmed that some of the covariates are associated with defaulters which are noted as the risk factors involved in defaulting.

When missingness was taken into consideration, the assumption of proportional hazard holds which confirms how fitted the model is for both response variables (defaulters and loss to follow-up), which makes this model to be preferred to the previous one without correcting for missingness. This shows that missingness has impact in this type of survival data since the significant covariates differ in both responses. We could not know the causes of patients default over the year.

7. RECOMMENDATION

In order to be able to know the causes of patients default, data on the ART (default) tracing visit should be provided. Data could then be analysed based on the true outcome status of patients at the ART tracing visit of Institute of Tropical Medicine, Antwerps. Through this data of ART tracing visit, the status of patients found to be alive, dead and possible reasons for unsuccessful tracing could be known. Easy assess to this information will go along way in tracing the causes of patients default. Also, having missing observations from some of the covariates most especially clinical stage with 39.33% shows that missingness might be difficult to avoid in this research work, therefore, the use of effective method of collecting data is important to minimize excessive missing observation.

8. REFERENCES

1. Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition Wiley Series in Probability and Statistics.

2. Allison, P. D. (1995). Survival *data Analysis using SAS: A practical Guide.*

3. Collett, D. (1994). *Modelling Survival data in Medical Research*. Chapman and Hall.

4. Hosmer, (Jnr), D.W., and Lemeshow, S. (1998). *Applied Survival Analysis, Regression modeling of Time to event Data*. 2nd Edition. Wiley Series in Probability and Statistics.

5. Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley Series in Probability and Statistics.

6. Kleinbaum, D. G. (1996). *Survival Analysis A Self-Learning Text.* New York: Springer-Verlag, Statistics in the Health Sciences.

7. Kleinbaum, D. G., and Klein, M. (2005). *Survival Analysis A Self-Learning Text* 2nd Edition. New York: Spring. Statistics for Biology and Health.

8. Klein, J.P., and Moeschberger, M. L. (1997). *Survival Analysis Techniques for censored and Truncated Data.* Statistics for Biology and Health. New York: Springer-Verlag.

9. Little, R.J.A., and Rubin,D.B.(1987). *Statistical Analysis with Missing Data*. New York: John Wilkey.

10. Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

11. Molenberghs, G., and Verbeke, G. (2007). *Longitudinal Data Analysis Course notes for International Study Programme* in Biostatistics, Universiteit Hasselt.

12. UNAIDS: *Global Summary of the HIV Epidemic in 2004*.

13. Vaida, F. (2007). *Survival Data Analysis Course notes for International Study Programme* in Biostatistics, Universiteit Hasselt.

14. Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

15. World Health Organization: (June 2005). *Progress on Global access to HIV Antiretroviral Therapy*: An update on 3 by 5.

16. Yu, J. K-L., Chen, S. C-C., Wang K-Y., Chang, C-S., Makombe S.D., Schouten, E.J., Harries, A.D., (July 2007).True outcomes for patients on antiretroviral therapy who are "lost to follow-up" in Malawi. Bulletin of the World Health Organisation, **85**, 7: 501-568.

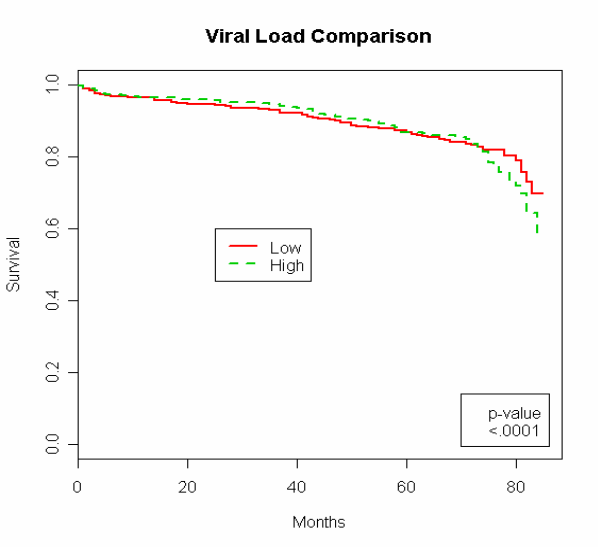
Websites

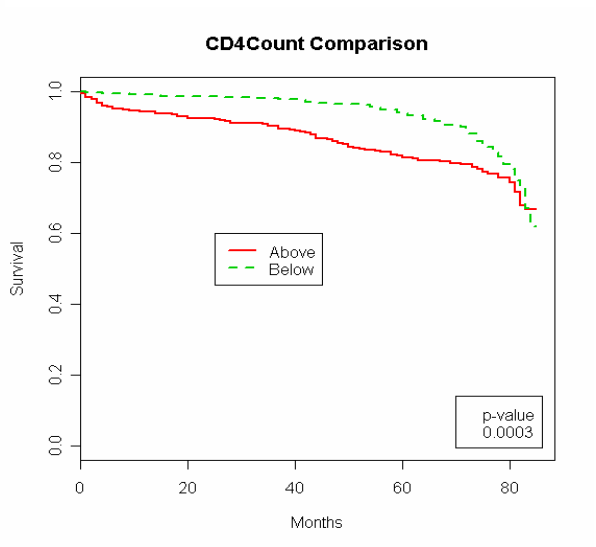
17. http://www.avert.org/hivstages.htm 28 April, 2007.

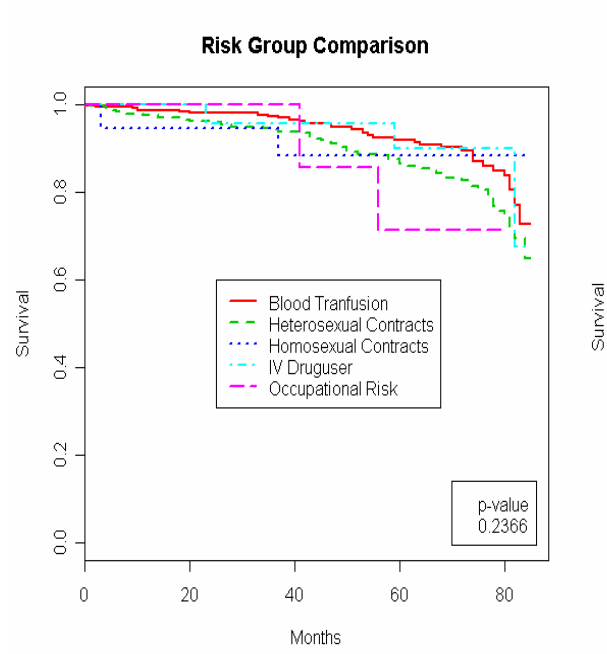
18. http://www.bradford.ac.uk/staff/ijhodgson/handouts/basicHIV.htm 01 May, 2007.

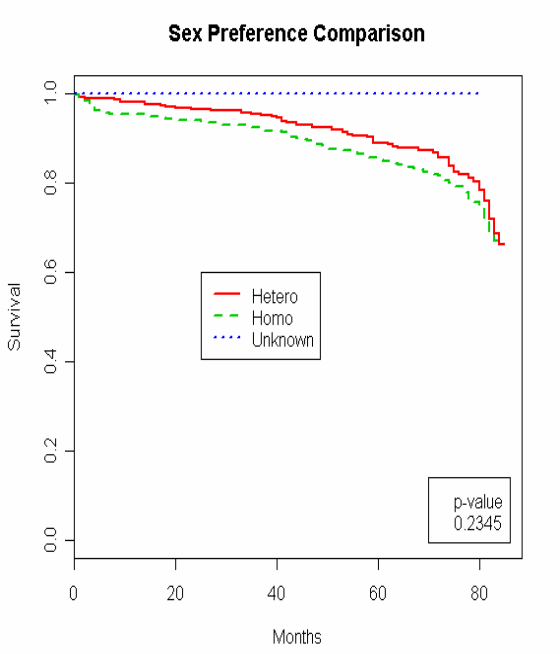
19. www.avert.org/introtrt.htm 01 May, 2007.

9. APPENDIX


Viral Load Comparison
Survival
Months
Low
High
p-value
<.0001


CD4Count Comparison
Survival
Months
Above
Below
p-value
0.0003


Risk Group Comparison
Survival
Months
Blood Tranfusion
Heterosexual Contracts
Homosexual Contracts
IV Druguser
Occupational Risk
p-value
0.2366


Sex Preference Comparison
Survival
Months
Hetero
Homo
Unknown
p-value
0.2345

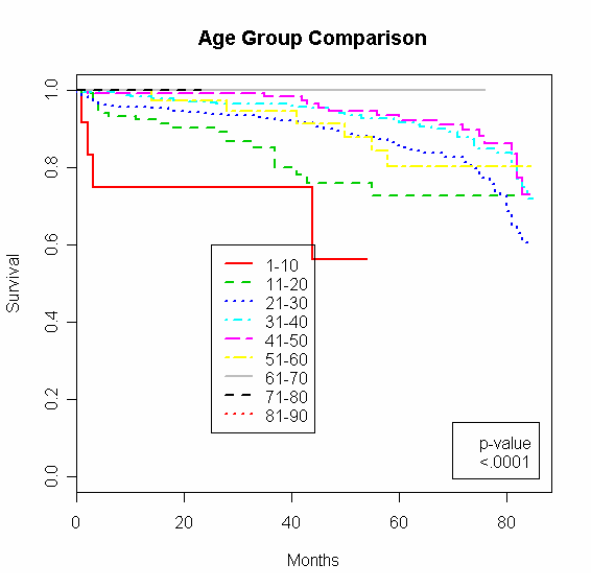


Figure 4.6: Graph of estimated survivorship function of Viral load, CD4 Count, Risk Group, Sex preference and Age group.

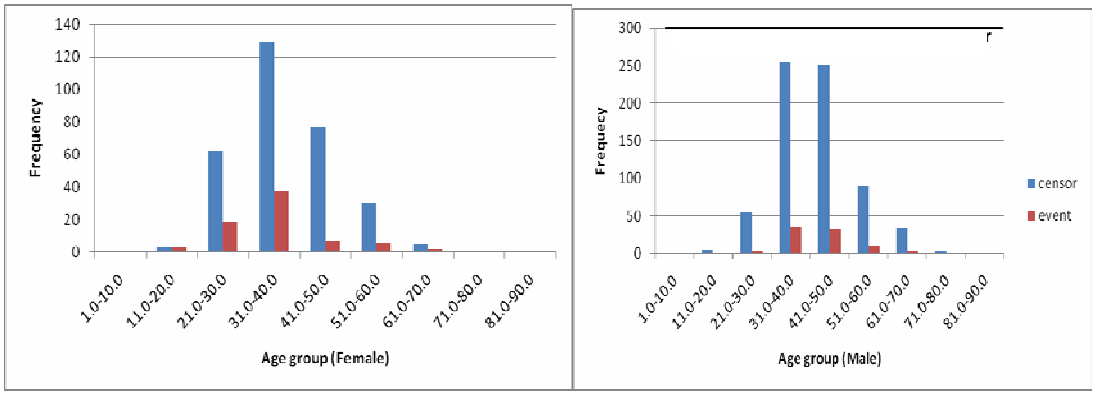
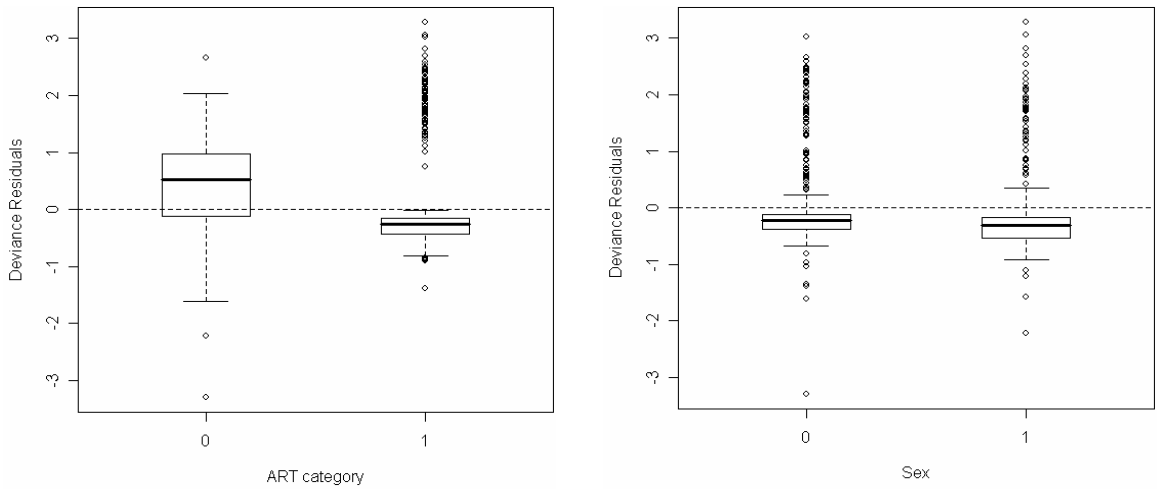


Figure 4.7: Graph of Defaulters by Gender and Age group



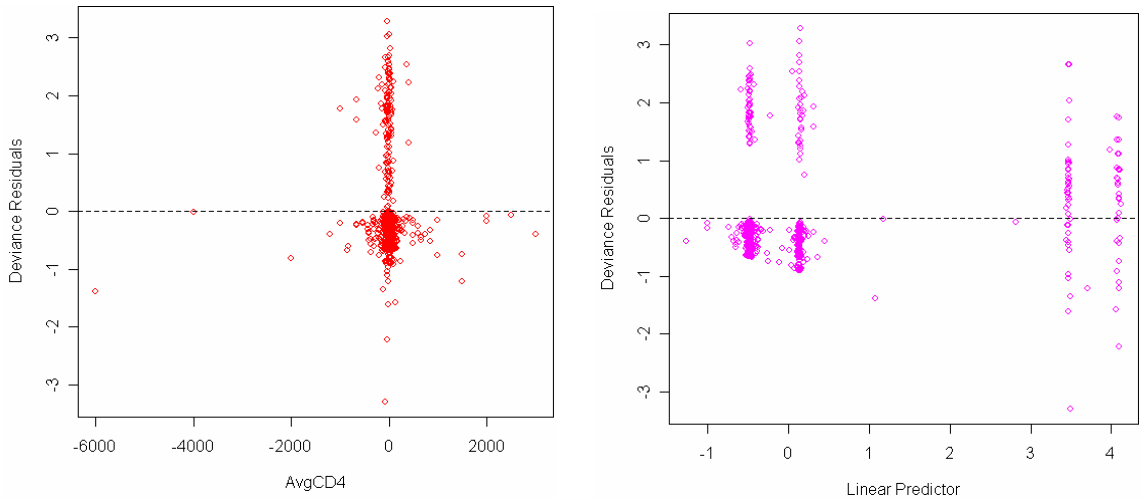


Figure 4.8: Plots of the deviance residuals against each predictor

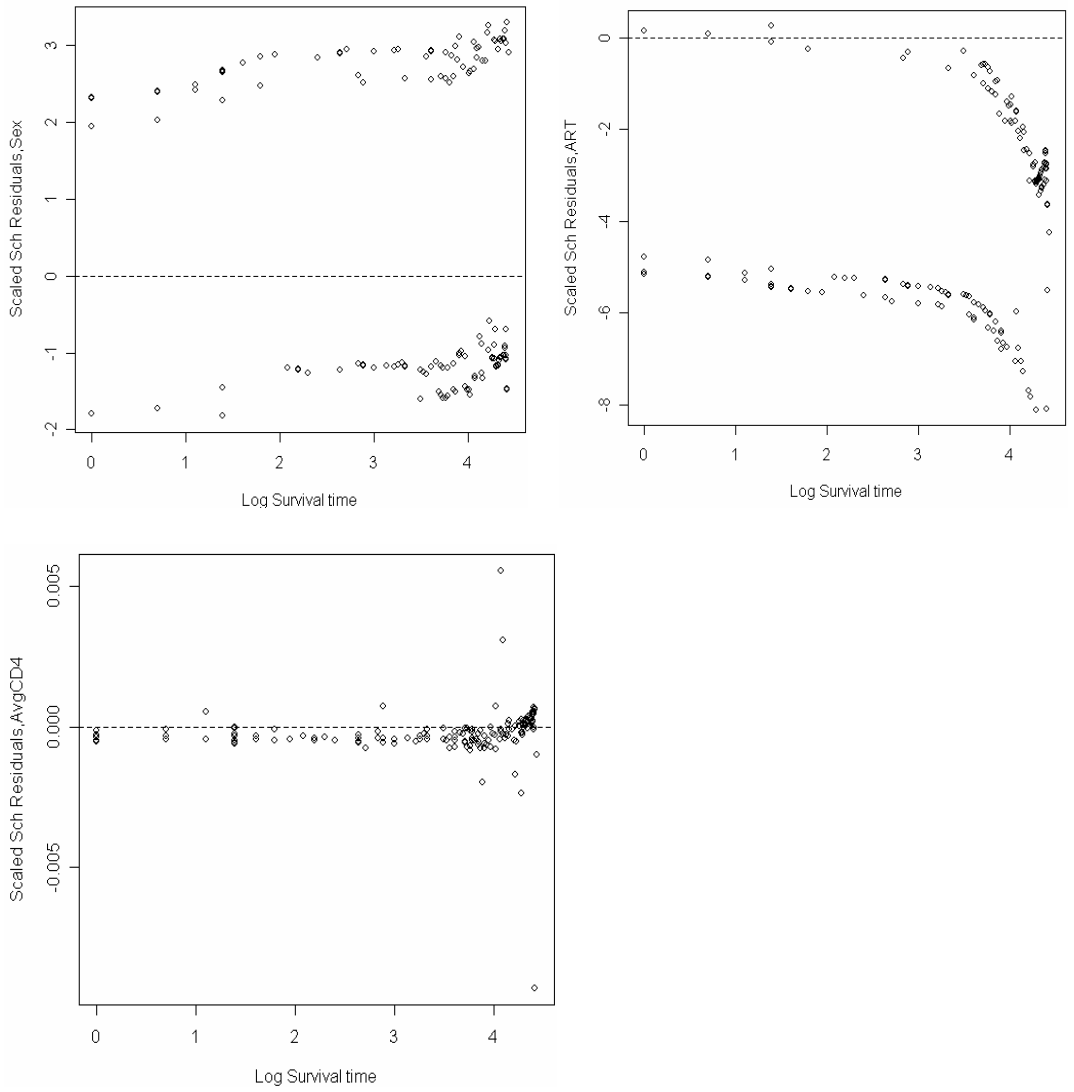


Figure 4.9: Plots of the weighted Schoenfeld residuals against log survival time

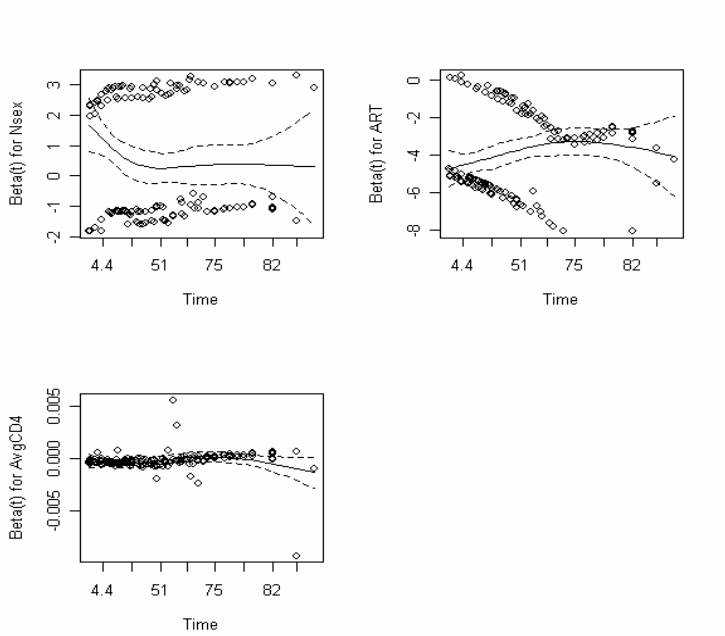


Figure 4.10: Plots of Schoenfeld Residuals vs. time

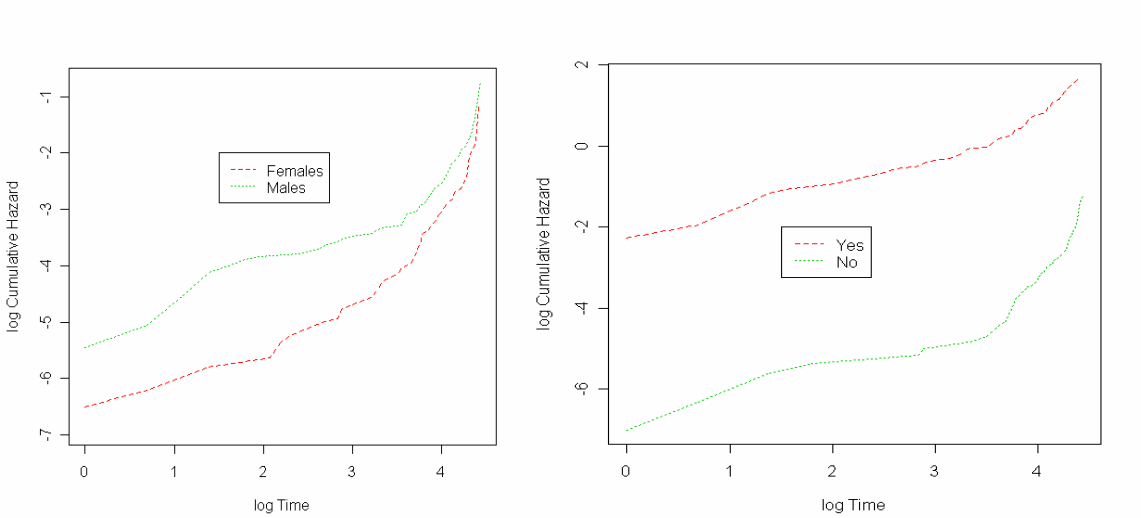


Figure 4.11: (a) Log [-log (survival)] plots for ART controlling for gender and Average CD4 count, (b) Log [-log (survival)] plots for gender controlling for ART and Average CD4 count.

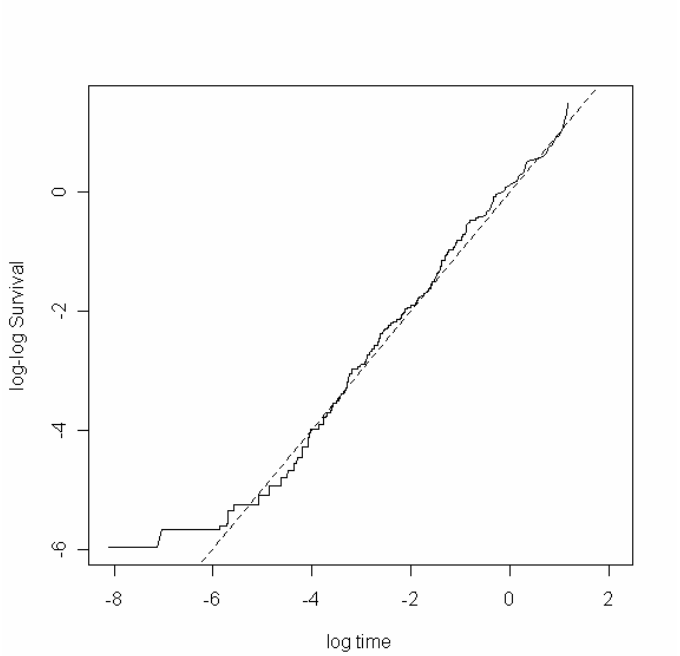


Figure 4.12: [-logS (t)] versus log (t) after correcting for missingness

Table 4.16: Table displaying Doctor and Patients information through the year.

S/N	Doctor ID	Event	Censored	No of Patients	Percent Defaulted	Duration of Years
1	5	4	72	76	5.26	2003-2006
2	392	19	100	119	15.97	2003-2006
3	978	31	162	193	16.06	2004-2006
4	1775	32	124	156	20.51	2004-2006
5	2094	25	161	186	13.44	2003-2006
6	3121	12	94	106	11.32	2003-2006
7	8452	3	13	16	18.75	2003-2006
8	15652	0	6	6	0.00	2003-2006
9	16977	7	54	61	11.48	2003-2006
10	17675	5	66	71	7.04	2003-2006
11	20369	7	37	44	15.91	2003-2006
12	26298	8	38	46	17.39	2003-2006
13	27399	6	24	30	20.00	2004-2006
14	27777	3	30	33	9.09	2004-2006
15	28981	1	16	17	5.88	2004-2006
16	31888	0	4	4	0.00	2006-2006
17	32403	0	3	3	0.00	2006-2006

Table 4.15: Summary of the Number of Event and Censored Values for all categorical variables

Categorical Covariates	Levels	Total	Event	Censored	Percent Censored	Percent Defaulted	Percent Missing
Gender	Male	783	87	696	88.89	11.11	
	Female	383	76	307	80.16	19.84	
	Total	1166	163	1003	86.02	13.98	1 (0.09%)
Risk Group	Blood transfusion	19	2	17	89.47	10.53	
	Heterosexual contacts	355	59	296	83.38	16.62	
	Homosexual contacts	334	36	298	89.22	10.78	
	IV druguser	28	3	25	89.29	10.71	
	Occupational risk	9	2	7	77.78	22.22	
	Total	745	102	643	86.31	13.69	422 (36.16%)
Clinical Stage	A: asymptomatic	103	14	89	86.41	13.59	
	B: symptomatic	416	69	347	83.41	16.59	
	C: aids	189	14	175	92.59	7.41	
	Total	708	97	611	86.30	13.70	459 (39.33%)
Sex Preference	Hetero	632	98	534	84.49	15.51	
	Homo	530	65	465	87.74	12.26	
	Unknown	1	0	1	100.00	0.00	
	Total	1163	163	1000	85.98	14.02	4 (0.34%)
Origin Group	0-199 (Europe)	596	69	527	88.42	11.58	
	200-299 (Asia)	31	4	27	87.10	12.90	
	300-399 (Africa)	279	39	240	86.02	13.98	
	400-499 (America)	25	6	19	76.00	24.00	
	>=500(Oceania)	17	4	13	76.47	23.53	
	Total	948	122	826	87.13	12.87	219 (18.77%)
Age Group	1-10	0	0	0	0.00	0.00	
	11-20	12	4	8	66.67	33.33	
	21-30	140	22	118	84.29	15.71	
	31-40	457	74	383	83.81	16.19	
	41-50	369	40	329	89.16	10.84	
	51-60	136	16	120	88.24	11.76	
	61-70	45	6	39	86.67	13.33	
	71-80	4	0	4	100.00	0.00	
	81-90	1	0	1	100.00	0.00	
	Total	1164	162	1002	86.08	13.92	3 (0.26%)
Viral Load	Low	485	84	401	82.68	17.83	
	High	682	79	603	88.42	9.20	
	Total	1167	163	1004	86.03	13.97	None
CD4Count	Above	645	115	530	82.17	17.83	
	Below	522	48	474	90.80	9.20	
	Total	1167	163	1004	86.03	13.97	None
ART	No	82	82	0	0.00	100.00	
	Yes	1085	81	1004	92.53	7.47	
	Total	1167	163	1004	86.03	13.97	None

Table 4.17: Result of the Cox's PH model and Stratified Cox's PH Model

Model Type	Variable	coef	exp (coef)	Se (coef)	z	p-value
Cox's PH Model	Sex	0.6231	1.8646	0.1599	3.89	9.8e-05
	ART	-3.9486	0.0193	0.1814	-21.77	0.0e+00
	AvgCD4	-0.0003	0.9997	0.0001	-2.10	3.6e-02
Stratified Cox's PH Model	Sex	0.5694	1.7700	0.1607	3.54	0.0004
	AvgCD4	-0.0002	1.0000	0.0001	-2.00	0.0450

Table 4.18: Test statistics for proportional hazards based on interactions between covariates and time

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Age	1	-0.0319	0.0229	1.9325	0.1645	0.969
AveragVL	1	-3.8309	2.9252	1.7151	0.1903	0.022
ART	1	-5.0814	0.7359	47.6734	<.0001	0.006
Sex	1	2.0007	0.5816	11.8351	0.0006	7.394
Aget	1	0.0010	0.0068	0.0218	0.8827	1.001
AverageVLt	1	1.0465	0.7157	3.5502	0.0695	2.851
ARTt	1	0.3659	0.2049	3.1856	0.0743	1.442
Sext	1	0.1339	0.1617	0.6906	0.6008	1.013

Table 4.19: Cross-Classification of Defaulters by Gender and Age

Sex	Age Group	Defaulters		Total	Gender	Defaulters		Total
		0	1			0	1	
Female	1-10	0	0	0	Male	0	0	0
	11-20	3	3	6		5	1	6
	21-30	62	19	81		55	3	58
	31-40	129	38	167		254	36	290
	41-50	77	7	84		252	33	285
	51-60	30	6	36		90	10	100
	61-70	5	2	7		34	4	38
	71-80	1	0	1		3	0	3
	81-90	0	0	0		1	0	1
Sub-Total		**307**	**375**		**Sub-Total**	**694**	**87**	
Total		**382**			**Total**	**781**		

Table 4.20: Final Model after Multiple imputation (Loss to follow-up)

Parameter	Estimate	Std. Error	95% Confidence Limits		DF	Pr > |t|
Clinical stage	-0.4613	0.2045	-0.8797	-0.0429	28.648	0.0319
Age	-0.0506	0.0126	-0.0753	-0.0259	13223	<. 0001
Sex	0.7171	0.2173	0.2909	1.1432	2945.9	0.0010

Table 4.21: Test statistics for PH based on interactions between covariates and time (Loss to follow-up)

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Clinicalstage	1	1.11447	1.51536	0.5409	0.4621	3.048
Age	1	0.01781	0.08284	0.0462	0.8298	1.018
Sex	1	2.69283	1.21788	4.8888	0.0270	14.773
Clinicalstaget	1	-0.61983	0.53139	1.3605	0.2434	0.538
Aget	1	-0.02187	0.02846	0.5901	0.4424	0.978
Sext	1	-0.71476	0.40774	3.0730	0.0796	0.489

Table 4.22: Formal test for the Proportional Hazard assumption (Loss to follow-up)

Variable	Chi square	P-value
Age	0.152	0.697
Sex	1.233	0.267
Clinical Stage	0.828	0.363
NA	2.170	0.652

Auteursrechterlijke overeenkomst

Opdat de Universiteit Hasselt uw eindverhandeling wereldwijd kan reproduceren, vertalen en distribueren is uw akkoord voor deze overeenkomst noodzakelijk. Gelieve de tijd te nemen om deze overeenkomst door te nemen, de gevraagde informatie in te vullen (en de overeenkomst te ondertekenen en af te geven).

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Defaulters in a cohort of HIV infected patients
Richting: **Master of science in Applied Statistics** Jaar: **2007**
in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Ik ga akkoord,

Oluwaseyi Akindunjoye

Datum: **27.08.2007**