A COX PROPORTIONAL HAZARDS MODEL FOR MID-POINT IMPUTED INTERVAL-CENSORED DATA

By

Arnold Rumosa Gwaze

A half-dissertation submitted in partial fulfilment of the requirements

for the degree of

MASTER OF SCIENCE IN BIOSTATISTICS AND

EPIDEMIOLOGY

Department of Statistics

School of Physical and Computational Studies

Faculty of Science and Agriculture

University of Fort Hare

Together in Excellence

Alice, South Africa

Supervisor: Prof J. Tyler

January 2011

Declaration

I, Arnold Rumosa Gwaze, hereby declare that this mini-dissertation submitted to the University of Fort Hare is my original work under the supervision of Prof. J. Tyler and has not been previously submitted to any university. Where reference to other researchers' work has been made and where assistance was rendered this has been duly acknowledged in the text.

Arnold Rumosa Gwaze _______________ **Date** _______________

SUPERVISOR

____

Prof. J. Tyler

Abstract

There has been an increasing interest in survival analysis with interval-censored data, where the event of interest (such as infection with a disease) is not observed exactly but only known to happen between two examination times. However, because so much research has been focused on right-censored data, so many statistical tests and techniques are available for right-censoring methods, hence interval-censoring methods are not as abundant as those for right-censored data.

In this study, right-censoring methods are used to fit a proportional hazards model to some interval-censored data. Transformation of the interval-censored observations was done using a method called mid-point imputation, a method which assumes that an event occurs at some mid-point of its recorded interval. Results obtained gave conservative regression estimates but a comparison with the conventional methods showed that the estimates were not significantly different. However, the censoring mechanism and interval lengths should be given serious consideration before deciding on using mid-point imputation on interval-censored data.

KEY WORDS: ACTG 181; interval-censored; Kaplan-Meier curve; logrank; mid-point imputation; proportional hazards; survival analysis.

Dedication

To my beloved wife, Elister and our two sisters Faith and Loreen.

Acknowledgements

Firstly, I would like to thank and to glorify God Almighty for giving me guidance throughout this study.

My utmost gratitude goes to all the members of the Statistics Department for the support they gave me. Special mention goes to my supervisor, Prof. J. Tyler, for patiently guiding, mentoring and inspiring me throughout this piece of work. Her expert advice did not fall on deaf ears. I acknowledge the financial support I received from the National Research Fund (NRF).

Acknowledgements and thanks are also due to Dr. Dianne Finkelstein (Harvard School of Public Health, Massachusetts General Hospital, Dana-Faber Cancer Institute, Boston, Massachusetts, U.S.A.) for her help and advice.

Lastly, special appreciation goes to my parents, family and friends for encouragement and support.

Preface

This dissertation is structured as follows:

- Chapter 1 comprises the introduction, research problem and research objectives.
- Chapter 2 focus on the overview of the literature review of nonparametric methods used in estimation of the hazard function when data is interval-censored. An overview of imputation is also given in this chapter.
- Chapter 3 gives the research methods and parameters used in this study.
- Chapter 4 reports on the results and analysis of the interval-censored data using the methods outlined in Chapter 3.
- Chapter 5 consists of a conclusion drawn from the methods suggested in Chapter 3. Recommendations and areas of future studies are also suggested in this chapter.

Table of Contents

List of Tables

List of Figures

Abbreviations

ANDA	Asymptotically Normal Data Augmentation
CI	confidence interval
CMV	cytomegalovirus
EM	expectation-maximization
LB	leverage bootstrap
MEMI	Multiple Edit / Multiple imputation
MLE	maximum likelihood estimator
MPS	maximum product spacings
NPMLE	nonparametric maximum likelihood estimator
PMDA	Poor Man’s Data Augmentation
PH	proportional hazards

Chapter 1 : Introduction

1.1 Background

Survival data, or time-to-event for humans, usually arise in biomedical studies when interest is focused on the time taken for a particular event to occur (Clark, *et al*, 2003). One of the most common sources of such data arises when time is recorded from some fixed starting point, such as surgery, to the death of the subject. In clinical studies, survival times often refer to time-to-death, development of a particular symptom or relapse after remission of the disease. Analysis of such data calls for special methods.

The first major reason why it is inappropriate to analyse survival data using the usual methods, such as the multiple regression techniques, t-tests or rank methods, is that residual survival times are usually not normally distributed (Altman, 1991). This condition violates the assumption for ordinary least squares multiple regression. Survival times normally follow an Exponential, Weibull or some other skewed distribution. Secondly, it is rarely feasible to observe the event of interest in all subjects. Such situations arise, for example, in longitudinal trials in which there is a periodic follow-up, or when the event of interest can only be determined by a laboratory test. In a comparative study to evaluate the effectiveness of two treatment regimens for breast cancer, for example, the event of interest may be time from diagnosis to death of the patient. If the time horizon of the follow-up is too short, such unobserved times are termed censored times indicating that the period of observation was cut off before the event of interest occurred (Altman, 1991).

1.2 Key functions in Survival Analysis

1.2.1 *Survival Function*

The survival or survivorship function is one of the basic quantities used to describe time-to-event phenomena. The survival function, denoted by $S(t)$, is the probability that an individual survives beyond time *t*, ie

$$S(t) = \Pr(T > t) \tag{1.1}$$

where T is the survival time. The cumulative distribution function, $F(t)$, given some probability density function, $f(t)$ of T, is a nonnegative function, with the area under $f(t)$ being equal to one, defined as;

$$F(t) = \Pr(T \leq t) \tag{1.2}$$

where $F(t) = \int_0^t f(x)dx$ for $t \geq 0$. Hence:

$$S(t) = 1 - F(t) \tag{1.3}$$

If the survival time T is a continuous random variable, the survival function, $S(t)$ and the probability density function, $f(t)$, have the following relationship;

$$f(t) = -\frac{d}{dt}[S(t)] \tag{1.4}$$

When T is a discrete, random variable, different techniques are required. Discrete, random variables in survival analyses arise due to rounding off measurements, grouping of failure times into intervals, or when lifetimes refer to an integral number of units. Suppose that T can take values $t_i, i = 1, 2, 3, ...$ with probability mass function defined by;

$$\Pr(t_i) = \Pr(T = t_i), i = 1, 2, 3, ... \tag{1.5}$$

where $t_1 < t_2 < t_3 < ...$

The Survival function for a discrete random variable T is thus given by;

$$S(t) = \Pr(T > t) = \sum_{t_i} p(t_i), i = 1, 2, 3, ... \tag{1.6}$$

Survival time can also be modelled using the hazard function.

1.2.2 *Hazard Function*

The hazard function is another fundamental basic quantity used in survival analysis. The hazard function, $\lambda(t)$ gives the instantaneous probability that an individual dies (experiences the event of interest) after time *t* given that the individual has survived and was alive at time *t*. It is defined as;

$$\lambda(t)=\begin{cases}\lim\limits_{\Delta t\to 0}\dfrac{P\left(t\le T<t+\Delta t\,|\,T\ge t\right)}{\Delta t} \textit{if} \quad T \quad \textit{is} \quad \textit{continous.}\\ P\left(T=t\,|\,T\ge t\right)=\dfrac{f(t)}{S(t-1)} \textit{if} \quad T \quad \textit{is} \quad \textit{discrete.}\end{cases} \tag{1.7}$$

If *T* is a continuous random variable, then, the hazard function, $\lambda(t)$, the survival function, $S(t)$ and the density function *f (t)*, are related as;

$$\lambda(t)=\frac{f(t)}{S(t)}=-\frac{d}{dt}[\log S(t)] \tag{1.8}$$

A related quantity is the cumulative hazard function, $\Lambda(t)$ which is defined as

$$\Lambda(t)=\int_0^t \lambda(x)dx=-\log[S(t)] \tag{1.9}$$

Hence for continuous lifetimes, the survival time is given by:

$$S(t) = \exp[-\Lambda(t)] = \exp[-\int_0^t \lambda(x)dx] \tag{1.10}$$

It can be observed that $\lambda(t)\Delta(t)$ may be viewed as the "approximate" probability of an individual of age t experiencing the event in the next instant. This function is particularly useful in determining the appropriate failure distributions utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time. There are various shapes for the hazard function; with some increasing, decreasing or being constant; others are bath-tub shaped, hump-shaped or possessing some other characteristic partly describing the failure mechanism. The only restriction on the hazard function is that it is always nonnegative, that is $\lambda(t) \geq 0$.

1.3 Some basic concepts in Survival Analysis

1.3.1 *Censoring*

The specific difficulties relating to survival analysis arise mainly from the fact that while some subjects experience the event of interest, others do not, during the period they are on the study. Hence survival times will be unknown for a subset of the study group. This phenomenon, censoring, may arise in any of the following ways:

- A patient has not yet experienced the relevant outcome, such as relapse or death, by the time of the close of the study, $T_i \leq t$, where T_i is the event time and t the time of death;
- A patient is lost to follow-up, moves away or dies from another disease, during the course of the study and the last time on record yields an incomplete waiting period;

- A patient experiences a different event that makes further follow-up impossible.

Censoring in biomedical studies exists in three forms; right-censoring, interval-censoring and left-censoring.

1.3.2 *Right Censoring*

If by the end of the observation period, the event of interest has not been observed, the time to event is said to be right censored. This type of censoring is most commonly encountered in biomedical research. Right censoring occurs when the lifetime (survival time) is known to exceed some specifiable value. Within right censoring, there are different types of censoring, with Type I and Type II mainly having to do with animal or industrial studies. In human studies, Type III censoring is the often encountered in clinical and epidemiological studies. Patients enter clinical trials at different times during the period of the study. For patients who die, where death is the event of interest, during the course of the study, their exact survival times are known. However, exact survival times for some patients may not be known. Possible reasons may be withdrawal from the study due to relocation of the patients, or to protocol specifications, due to some adverse effects. Other patients may just get lost-to-follow-up, whilst others may still be alive up to the end of the study. For patients whose survival times cannot be known, their survival times are at least the period from their entrance into the trial to the last contact or observation time. Patients still alive till the end of the study have survival times that are at least from entry to the end of the study and they are censored at the end of the study (Lee & Wang, 2003).

Let T_i be the event time for the ith subject whose event time lies in the interval (t,∞) but also exceeds the study period. Let the censored time be C_i, then the observed time for the subject is given by $\min(T_i,C_i)$. The data are usually represented as (T_i,δ_i) where T_i is the recorded time and δ_i is the censoring indicator variable defined as:

$$\delta_i = \begin{cases} 1 & if \quad T_i = t_i \\ 0 & if \quad T_i > t_i \end{cases} \tag{1.11}$$

Hence $\delta_i = 1$ for exact times and $\delta_i = 0$ for right-censored times. For the illustration in Figure 1.1 (Page 7), $\delta_i = 0$ and the survival time is right-censored.

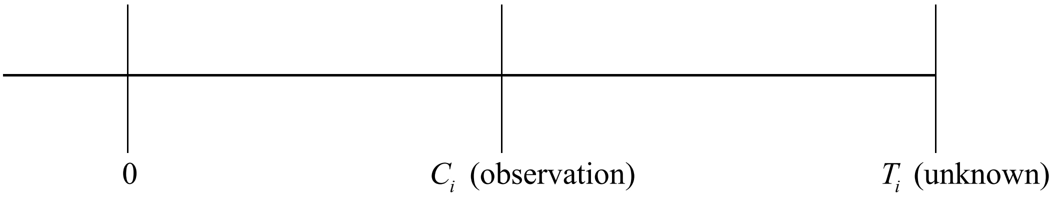


Figure 1.1 : Example of right censored data

1.3.3 *Interval Censoring*

Censoring is not only confined to right-censoring; when some event of interest can only be detected by a laboratory test, the event is thus likely to occur between two tests, hence the exact time of occurrence may not be observed. All information known is that the occurrence occurred within a known time interval (see Figure 1.2, page 8). This partial knowledge gives rise to interval-censored observations.

Failure-time data are said to be interval-censored when the event of interest is not observed exactly but instead, is only known to lie in an interval. Such data usually occur in clinical or longitudinal studies. In such studies, some failure events can only be detected by a screening examination. For example, Kim, *et al* (1993) describe data from an HIV screening study in which stored blood samples were tested retrospectively to determine seropositivity. For patients who had seroconverted, it is known that this change occurred between the time of their last negative screening and the time of first positive screening. Interval-censored data also appear in clinical trials, for example there are pre-scheduled periodic follow-ups, for example weekly for clinically observable change or response. Some subjects may miss visits for a few weeks and then return with a changed state. Remission duration is one of the most important clinical variables for which interval censoring occurs; the time and degree of remission and relapse are frequently unknown. When each patient only has one examination time, the data is termed current status.

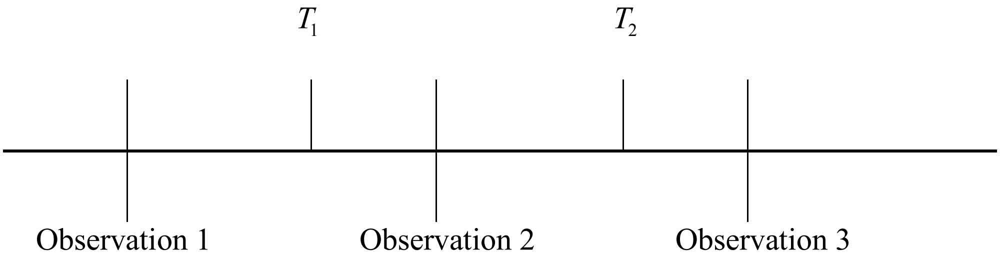


Figure 1.2 : Example of interval censored data

Another special type of interval censoring, called left censoring, occurs when the event of interest has already occurred by the time of the first observation.

1.3.4 *Left Censoring*

Left censoring is a special type of interval censoring that occurs when all information known about an observation on a variable T is that it is less than some known value (Allison, 1995). It occurs when a sample is observed at a time when some of the individuals have already experienced the event of interest. In an observational study of the onset of menstruation of 12-year-old girls, the event of interest is menstruation. At this age some girls might already have begun menstruating. For these girls, their age at menarche is left-censored at age 12.

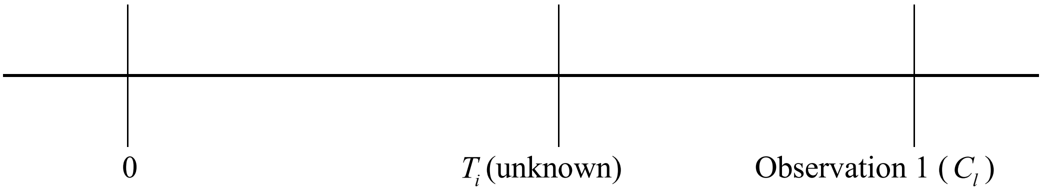


Figure 1.3 : Example of Left-censored data

1.3.5 *Truncation*

Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (X_L, X_R) are observed. An individual whose event time is not in this interval is not observed and no information on this subject and no information on this subject is available to the investigator. This is in contrast to censoring where there is at least partial information on the subject. Inference for truncated data is restricted to conditional estimation.

When X_R is infinite, then we have left truncation. We only observe those individuals whose event time T exceeds the truncation time X_L, thus T if and only if $X_L < T$.

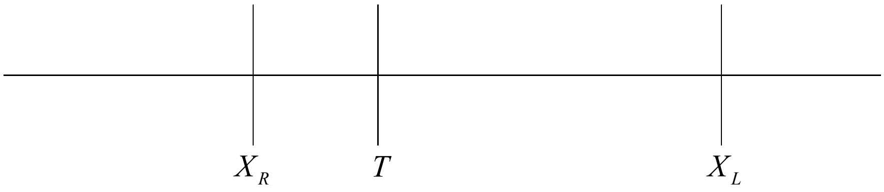


Figure 1.4 : Example of Truncated data

1.3.6 *Cox Proportional Hazards Model*

The Cox proportional hazards model possesses the property that different individuals have hazard functions that are proportional and independent of time. That means that for two subjects with prognostic factors $X_1 = (X_{11}, X_{21}, ..., X_{p1})'$, and $X_1 = (X_{12}, X_{22}, ..., X_{p2})'$, then the proportion of their hazards, their hazard ratio, $\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t)\exp\{(X_{11}, X_{21}, ..., X_{p1})'}{\lambda_0(t)\exp\{(X_{12}, X_{22}, ..., X_{p2})'}$ is dependent on the prognostic factors only. The model does not require knowledge of the underlying distribution. The hazard function can take on any form. More about the model is given in Section 3.3.3.

1.4 Key requirements for the analysis of survival data

Five major conditions are pointed out by Clark, *et al* (2003) as being of importance for consideration in the analysis of survival data. These conditions are uninformative censoring, length of follow-up, completeness of follow-up, cohort effect on survival and between-centre differences.

1.4.1 *Uninformative censoring*

Standard methods used to analyse survival data with censored observations are valid only if the censoring is non-informative (Lagakos, 1979). This condition means that censoring carries no prognostic information about subsequent survival experience; that is, subjects who are censored because of loss to follow-up at any time should be as likely to experience the event as those who remain in the study. In contrast, informative censoring may occur when subjects withdraw from a clinical trial because of drug toxicity or worsening clinical condition. Standard methods for survival analysis are not valid when there is informative censoring.

1.4.2 *Length of follow-up*

Analysis of results from a study is influenced by the design of the particular trial. Time-to-event studies must have sufficient follow-up to capture enough events and thereby ensure there is sufficient power to perform appropriate statistical tests. The proposed length of follow-up for a prospective study is based primarily on the severity of the disease or prognosis of the subjects and the clinical relevance of the observed end-points. For example, for a lung cancer trial, a 5-

year follow-up might be more than adequate, but this follow-up duration will only give a short-to-medium-term indication of survival among breast cancer patients. This is due to the fact that lung cancer is a more dynamic disease than breast cancer. More can be learnt in five years from lung cancer patients in comparison with breast cancer patients.

1.4.3 *Completeness of follow-up*

Every uncensored subject should be included in the analysis until they are censored, but completeness of follow-up is still of importance. Unequal follow-up between different groups, such as treatment arms, may introduce bias in the analysis. Generally, unequal follow-ups caused by differential drop-outs between arms of a trial in a cohort study are worthy of investigation.

1.4.4 *Cohort effect on survival*

The assumption of homogeneity of treatment and other factors during the follow-up period plays a pivotal role in the validity of results by preventing introduction of bias. In a long-term observational study of cancer patients, the case mix may, however, change over the period of recruitment, or there may be an innovation in ancillary treatment. The Kaplan-Meier method assumes that the survival probabilities are the same for subjects recruited early and late in the study. On average, subjects with longer survival times would have been diagnosed before those with shorter times, and changes in treatments, earlier diagnosis or some other change over time may lead to spurious results. The assumption may be tested, provided that there is enough data to estimate survival probabilities in different subsets of the data and, if necessary, adjusted for further analyses.

1.4.5 *Between-centre differences*

In a multicentre study, there is a danger of having inconsistencies which may introduce bias in the results. For instance, diagnostic instruments, such as staging classification and treatments should be identical. Differences between prevailing prognostic factors at different study centres should be adjusted for in an analysis.

1.5 Research Problem

The development of the Cox proportional hazards regression model (Cox, 1972), marked a milestone in the analysis of right-censored survival failure time data. This model was a refinement of earlier works (Mantel & Haenszel, 1959; Peto & Peto, 1972). The advantages of the method are that it is distribution-free and results in an estimate for the risk for failure associated with a vector of covariates. The proportional hazards model has been extensively developed and applied right-censored data (Kalbfleisch & Prentice, 1980; Cox & Oakes, 1984).

Whereas attention has been focused on right-censored survival times, interval-censored data have become more common because of the increased use of laboratory measures to monitor progression of chronic diseases such as AIDS and cancer. Decisions on the continuation or termination of trial progress are now made on the basis of lab tests. Researchers have come up with many methods for right-censored estimation and only a few of the now highly sought-after interval-censored methods. Thus, even software for interval-censored data analysis methods are not yet comprehensive (Lesaffre, *et al*, 2005). Hence to analyse interval-censored data, one can make use of the few methods available or alternatively convert the data into right-censored data and analyse the data using the usual right-censored methods.

Mid-point imputation is one attractive way of applying the right-censored methods on interval-censored data. In this scheme, the event of interest is assumed to have occured at the mid-point of the interval. The problem to be investigated is whether the use of such mid-point imputation on interval-censored data can yield comparable results against the interval-censored methods. A proportional hazards model will be fitted to some ACTG 181 data set.

1.6 Objectives

The aim of this study was to apply right-censored methods to interval-censored data and to compare the results with those obtained using interval-censored methods. The objectives of the study are:

(i) to fit the right-censored Cox Proportional Hazards model to some ACTG 181 interval-censored data set using the mid-point imputation method;

(ii) to compare the mid-point imputation results from this study with results from other methods.

Chapter 2 : Literature Review

2.1 Introduction

In survival or time-to-event analysis, failure time can be defined as the time-to-occurrence of some event of interest. Examples of survival times include the duration of a patient from the moment of some infection to appearance of symptoms related to the infection, time to death of an HIV subject from the onset of some HIV treatment, and so on. One feature of survival data is that the data is often incomplete. This happens when the endpoint of the experiment is not seen to occur. In a clinical trial, incomplete observation of failure time may be due to lost-to-follow-up of subjects, death due to other causes or due to the expiration of the study. For example, it is common that not all subjects survive a trial to experience the event of interest. Such incomplete observation of failure times is called censoring (Cox & Oakes, 1984).

2.2 Types of censoring

Survival times can be left-, right- or interval-censored depending on the censoring mechanism. Suppose each subject is examined a number of times for some event of interest in a clinical trial. Left-censoring occurs when the subject has already experienced the event at the very first examination. Right-censoring occurs when a subject has not experienced the event of interest by the time of the last examination; the subject may be lost to follow-up or relocated without leaving a forwarding address, died from an unrelated cause or the trial may be terminated before the event has occurred. Interval censoring arises naturally when the response times are obtained from a clinical trial or a longitudinal study in which there is a periodic follow-up. An individual who is monitored weekly for a clinically observable change or response may miss visits for a few weeks, and return in a changed response state. In such a case the subject is said to be interval-censored. If all subjects were to keep their appointments, and get visited or visit the clinic as pre-arranged, the data would either be exact or right-censored. Examples of interval-censored data usually arise from research on HIV and AIDS, because important events such as infection, seroconversion, and extent of disease progression, which are measured by the steadily decreasing concentration of CD4+ cell counts, are ascertainable only by laboratory tests and do not produce uniquely identifiable clinical symptoms. An example is given by a follow-up study on HIV-negative persons who are at high risk for becoming HIV infected; such studies are conducted in preparation for testing HIV vaccines (Hoff, 1994).

A special case of interval-censored data is current-status data, where individuals are each examined only once after enrolment. In such studies, the main aim is the determination of the distribution of some variable in relation to some disease or life event.

In this chapter, a brief overview of examples of interval-censored data is presented, followed by a review of the nonparametric methods developed for interval-censored failure times.

2.3 Examples of Interval-censored data

2.3.1 *Haemophilia data*

A multi-centre prospective study was conducted in 1980's to investigate HIV-1 infection rate among people with haemophilia (Kroner, *et al*, 1994). The subjects were at risk of HIV-1 infection from blood products such as factor VIII and factor IX made from donors' plasma. In this study, interval-censored data were observed for subjects' HIV-1 infection times. The subjects were categorized into one of four groups according to the average annual dose of the blood products they received: high-, medium-, low-, or zero-dose group. The goal of this study was to compare the HIV-1 infection rates between treatment groups. More details about this study can be found in (Kroner, *et al*, 1994).

2.3.2 *Breast cancer (cosmesis) data*

Breast cancer data (Goggins & Finkelstein, 2000; Finkelstein, *et al*, 2002) reports 94 early breast cancer subjects or subjects in two treatment groups, radiotherapy alone and radiation therapy together with adjuvant chemotherapy. Among the subjects, 46 received radiotherapy and 48 received radiation therapy and adjuvant chemotherapy. In this study, subjects were examined periodically. Examination times differed from subject to subject as some of them missed their visits. One objective of this study was to detect whether chemotherapy changes the rate of deteriorations of the cosmetic state. Breast retraction, a response that has a negative impact on the overall cosmesis appearance, was taken to be the event of interest which led to interval-censored data. The data are presented in Appendix A on page 76. References that discuss this data set include (Goggins & Finkelstein, 2000; and Finkelstein, *et al*, 2002).

2.3.3 *ACTG 181 data*

The ACTG 181 data come from an AIDS observational study conducted by the AIDS Clinical Trials Group (ACTG). More details of the data set is given in Section 3.2.1.

2.4 Nonparametric Methods of estimation

Goggins & Finkelstein (2000) focused on the methodology developed for analysing a multivariate interval-censored data set from an AIDS observational study, the ACTG 181. The methodology developed was based on the discrete proportional hazards model (Prentice & Gloeckler, 1978). Goggins & Finkelstein, 2000 suppose K types of failure monitored on N subjects with the time axis divided into m time intervals. If data were completely recorded for all subjects, then the result would be K contingency tables each with time along the columns and Z on the rows. Let the probability of a person i with $z=0$ experiencing the kth failure in the ith interval be denoted by $g_{i,k}$. Then under the proportional hazards (PH) model proposed by Finkelstein (1986), the probability that a person with covariate z is free of the kth failure for more than r periods is given by $\left(\sum_{i=r+1}^{m} g_{i,k}\right)^{\exp(\beta_i z)}$ which reduces to the (Cox, 1972) proportional hazards model as the number of groups get large. Goggins & Finkelstein (2000) then performed Monte Carlo simulations with bivariate dependent failure times generated from Gumbel (1960) exponential distribution whose results indicate some bias in the estimates of the β_i s. The use of naive estimates resulted in underestimation of the standard error. The new methodology converged for data set of size $n = 200$. The method proved to be applicable to covariates measured at a single point in time, as well as to time-varying covariates as long as the covariates are ancillary or external. However, Goggins & Finkelstein (2000) noted that the method could be problematic for computational reasons if the number of parameters was close to the total sample size, as noted earlier on by Finkelstein (1986). The assumption made was that the mechanism that produces the interval censoring is independent of the failure process.

Peng & Dear (2000) studied a general nonparametric mixture model and applied it to the censored data generated from the cure rates on breast cancer subjects after some simulation studies. Peng & Dear (2000) built their model on the model by Kuk & Chen (1992) by including the expectation-maximization (EM) algorithm, marginal likelihood approach and multiple imputation. The model also extends proportional hazards model by allowing some of event-free subjects and investigating covariate effects on that group (Cox, 1972). The PH assumption was employed in the analysis of covariates effects on failure times of cured subjects. The estimation is a combination of marginal likelihood approach for the Cox PH model and the EM algorithm. Simulations studies of sample size 500 produced comparable results against the Kuk & Chen (1992) logistic regression with the PH model. For no covariate considered for failure time of uncured subjects, the model reduces to the one proposed by Taylor (1995) and when there is no cure fraction, the model reduces to the Cox PH model. The multiple imputation method was employed to estimate the observed information matrix of regression parameters for the failure time of uncured subjects.

In 2001, Fang & Sun (2001) examined at the consistency of nonparametric maximum likelihood estimation of survival time function on doubly interval-censored univariate survival data. Doubly censored survival times arise when the initial and subsequent event times are both interval-censored. For example when some infection causing some disease, the initial event and the onset, the subsequent event, of the particular disease cannot both be determined exactly but are known to lie in some intervals. The doubly interval-censored time is the failure time of interest defined as the lapse between the initial event and a subsequent event and observation on both events are interval-censored. The estimator proposed by Fang & Sun (2001) is recommended because it uses a full likelihood and also that it uses a 2-step procedure. Nonparametric estimation on such data has been done using AIDS studies (DeGruttola & Lagakos, 1989) and later propositions have been reviewed and improved, including asymptotic properties (Gomez & Lagakos, 1994; Gomez & Calle, 1999).

Recently, in 2009, Deng & Fang (2009) went ahead to extend the work of Fang & Sun (2001) by studying the asymptotics for nonparametric likelihood estimation of multivariate doubly-censored data. Such data arise in health-care field. An example the longitudinal prospective oral health study was done by Komárek & Lesaffre (2006) and Komárek & Lesaffre (2008) in Flanders, Belgium. In the study, children born in 1989 were examined annually with the primary interest being to investigate the influence of sound versus affected deciduous second molars on the caries susceptibility of the adjacent permanent first molars. The onset time, $U_{i,l}(l=1,...,4)$ is the age of the i^{th} child at which the l^{th} permanent first molar emerged. The failure time, $V_{i,l}$, is the onset of caries of the l^{th} permanent first molar. The survival time, $T_{i,l}$, was the lapse of time from tooth emergence to the onset of caries. Since both the time to tooth emergence and the onset of caries are only known to lie within about one year, $T_{i,l}$ is multivariate doubly interval-censored. Using bivariate interval-censored mechanisms, Deng & Fang (2009) observed that the bigger the number of censoring interval is, the faster the convergence rate is. This pattern arises because the bigger the number of intervals, the more information becomes available for analysis.

Cai & Betensky (2003) introduced an approach for estimating the hazard function for interval-censored and right-censored survival data. Cai & Betensky (2003) weakly parameterized the log-hazard function with a piecewise-linear spline and provided a smooth estimate of the hazard function by maximizing the penalized likelihood through a mixed model-based approach. They argued that the Cox PH model works well for the left-censored and the right-censored failure times, but the situation becomes complex when it comes to interval-censored data. For the interval-censored data, estimation of the regression parameter, β, cannot be easily separated from estimation of the baseline hazard function. This approach was built upon earlier work on nonparametric models of weakly parameterizing the hazard function (Whittemore & Keller, 1986; Rosenberg, *et al*, 1994; Kooperberg, *et al*, 1995). Cai & Betensky (2003) assumed a log-linear spline mixed model for the baseline hazard function, and the Cox PH model for the covariate effect. Cai & Betensky (2003) showed that with the penalized quasi-likelihood (PQL), using approximation, the estimate (Breslow & Clayton, 1993) is equivalent to the penalized spline fit with a quadratic penalty on the knot coefficients. Cai & Betensky (2003) recommended the method to be usually fast, relatively simple to program, and it has the advantage of simultaneously calculating the regression parameter and the hazard function. However Cai &

Betensky (2003) leave the associated variability in calculating the regression parameter for future research even though they claim the variability to be negligible for large samples.

Betensky, *et al* (2002) proposed a local likelihood-based nonparametric method for estimating the hazard function. They illustrated their method on two sets of data; the breast cosmesis data and the HIV-1 infection rates among haemophiliacs. The EM algorithm described by Betensky, *et al* (1999) where no covariates are considered, is extended in this study. Betensky, *et al* (2002) describe a four-step algorithm where Step 0 is called the Initialization step. In Step 1, the hazard function is estimated, assuming that the covariates are known. This step, which has two substeps is iterative and had to be terminated when convergence is reached. Step 3, is a repetition of Steps 1 and 2 until convergence is reached as well. The proponents of this method claimed, after reanalysing the breast-cosmesis and haemophilia data sets, that making weak assumptions on the baseline hazard and coming up with a smooth baseline hazard increases the interpretability and understanding of the failure process. This outcome, according to Betensky, *et al* (2002), tends to be useful when dealing with multiple covariates. Although their approach has the benefit of being derived from the local likelihood function, it requires manual entry of a bandwidth parameter that determines the amount of smoothing for the hazard function estimate. Furthermore, the analytic standard errors were not derived, necessitating the use of bootstrap, which tends to be computationally extensive.

A class of procedures for local likelihood estimation from data that are either interval-censored or that have been aggregated into bins was proposed in 2005 (Braun, *et al*, 2005). One such procedure relies on an algorithm that generalizes existing self-consistency algorithms by introducing kernel smoothing at each step of the iteration. The entire class of procedures yields estimates that are obtained as solutions of fixed point equations. Kernel density estimation tends to have an appealing interpretive basis. Central to its use are kernel weights which depend on the proximity of an observation to the point of estimation, lending the estimator a local interpretation. As for interval-censored data, an observation is known to lie within some interval and it seems natural to define the weight as the conditional expectation of the kernel over that interval. Doing so yields an estimator that retains the interpretive appeal of a kernel density estimate. When the conditional expectation is computed with respect to the density estimate, a fixed point equation arises. Solving the equation iteratively leads to a generalization of the classical self-consistency algorithms of Efron (1967); Turnbull (1976) and Li, *et al (*1997). The estimator avoids some arbitrary aspects associated with the standard technique of directly smoothing the NPMLE of the cumulative distribution function. Braun, *et al* (2005) commented that the approach proposed, recasting a local expectation-maximization (EM) algorithm as Newton iteration, permitted some formal developments concerning convergence.

Hudgens, *et al* (2001) derived the NPMLE of the cumulative incidence functions for competing risks survival data subject to interval censoring and truncation. The method was illustrated on the Bangkok Metropolitan Administration of Thailand injection drug users (BMA IDU) cohort established in 1995 to assess the feasibility of some phase III HIV trial (Vanichseni, *et al*, 2001). Competing risk data arise when the event of interest can be achieved through more than one route. This possibility is where it is of interest to determine the hazard rate of a certain type of failure amongst a number of types. The BMA IDU cohort study was designed to measure rates of successful follow-ups on HIV incidence and to assess the effectiveness of some HIV prevention measures. Subjects in the study were monitored for HIV seroconversion. Some seroconverted to subtype B while others seroconverted to subtype E. It was assumed that each subject would seroconvert to only one subtype and that the infections were mutually exclusive in the subjects. Hence the subjects were subject to competing risks between the two subtypes. Since the cumulative incidence function NPMLEs give rise to an estimate of survival distribution which can be undefined over a potentially larger set of regions than the NPMLE of the survival function obtained ignoring failure type, they considered an alternative pseudolikelihood estimator. In the competing risks setting, the cumulative incidence function was estimated. The NPMLE of the cumulative incidence function for right-censored, competing risks survival data is given in Kalbfleisch & Prentice (1980). In the absence of competing risks, Peto (1973) first characterised the survival function NPMLE for interval-censored failure time data and used a constrained Newton-Raphson algorithm for estimation. Turnbull (1976) extended the work of Peto (1973) to allow for truncation while using a self-consistent algorithm. Hudgens, *et al* (2001) recommended that further studies be carried out on their method so as to investigate consistency, rates of convergence and asymptotic distributions.

Gentleman & Vandal (2001) used graph theory to present methods for finding the NPMLE of survival time distributions. Gentleman & Vandal (2001) used the intersection of graphs to simplify the problem. Gentleman & Vandal (2001) showed that right-, interval- or double-censored or current status data can be represented in terms of the intersection of their graphs. Combinatorial algorithms can be used to find the important structures, the maximal cliques. The algorithms can be extended to deal with bivariate data and there are no fundamental problems extending the methods to higher dimensional data. The study shows how to obtain the NPMLE using convex optimization methods and methods for mixing distributions. The implementation of these methods is greatly simplified through the graph-theoretic representation of the data. One drawback for the method is that it fails on the uniqueness of solutions (Gentleman & Vandal, 2001). Gentleman & Vandal (2001) discovered that the algorithms mentioned could be too slow, might not find zeros or always converge. Gentleman & Vandal (2001) recommended further research, involving comprehensive comparisons to determine which method should be preferred.

2.5 Imputation

Recently Zhang, *et al* (2009) define imputation in relation to HIV infection. They let HIV infection be the origin event and death be the endpoint event. If X_i denotes the infection time for subject $i = 1,...,n$ and assume that X_i is interval-censored, that $L_i < X_i \leq R_i$, where L_i and R_i are known pre-scheduled times, then Zhang, *et al* (2009) classify imputation methods into two categories; simple imputation and probability-based imputation. Simple imputation methods include the right, mid-point and left imputations. Probability-based imputation requires estimating the distribution for HIV infection time based on observed intervals. Such estimates have been studied by many authors (Grooeneboom & Wellner, 1992; Dempster, *et al*, 1977; Turnbull, 1976). Zhang, *et al* (2009) also suggest the use of the conditional mean, conditional median, conditional mode, multiple imputation and random imputation under the probability-based imputation. Simulation studies showed that the right imputation does not perform well in estimating the Kaplan-Meier curve in the one-sample case with the mean and median imputations preferable in the two-sample case. It was generally observed that in all cases, as the interval width decreases, the performance of each imputation method improves.

Multiple imputation (MI) was first proposed by Rubin in the 1970's as a possible solution to the problem of survey non-response (Rubin, 1977; Rubin, 1978). Rubin emphasized orderliness in handling missing data. In their research entitled “Inference Based on Imputed Failure Times for the Proportional Hazards Model With Interval-Censored Data”, Satten, *et al* (1998) proposed an approach to the proportional hazards model for interval-censored data. Parameter estimates were obtained by solving estimating equations that are the partial likelihood score equations for the full-data proportional hazards model, averaged over all rankings of imputed failure times consistent with the observed censoring intervals. Imputed failure times are generated with the proportional hazards regression parameters. The method is seen to work well, through using simulation studies; even when the baseline parametric form is misspecified, an improvement to the method of Satten, (1996), especially in cases of extreme censoring. The estimating equations are solved using the MC techniques. Satten, *et al* (1998) presented a recursive stochastic approximation scheme that converges to the zero of the estimating equations. The solution has a random error that is asymptotically normally distributed with a variance-covariance matrix that can itself be estimated recursively. The simulation studies also confirm that the proposed estimator provides an advantage over a fully parametric estimator in that dependence of the estimates on correct specification of the baseline distribution is reduced. The other advantage of the proposed method to the missing-rank approach of Satten (1996) is that generating the imputed failure times is much easier, more efficient and faster than the rank-generating scheme used in the missing-rank method, which requires a Gibbs sampler. Also when data-sets are heavily censored, the missing-rank method seems to require larger sample sizes to produce an unbiased estimate. Having an imputed failure time available allows straightforward generalizations of the interval-censoring problems.

Pan (2000) proposed a general semiparametric method based on multiple imputation for Cox regression with interval-censored data. The method consists of iterating the following two steps; firstly, the finite interval-censored data, exact failure times are imputed using any of the two schemes outlined by Wei & Tanner (1991), the PMDA or the ANDA, based on the current estimates of the regression coefficient and the baseline survival curve. Secondly, a standard statistical procedure for right-censored data, such as the Cox partial likelihood method, is applied to imputed data to update the estimates. Pan (2000) reported that the method is easy to implement and can take full advantage of existing techniques for right-censored data. Through simulation, Pan (2000) reported that the method performs better than the NPMLE does in estimating the regression coefficient in the Cox proportional hazards model with small to medium samples. Pan (2000) confirms that the Poor Man's Data Augmentation (PMDA) works reasonably well in many situations but may underestimate the true variability if the degree of missingness is severe. The performance of the ANDA is satisfactory in all simulation setups. Hence the ANDA is recommended.

Faucett, *et al* (2002) developed an approach, based on MI, to using auxiliary variables to recover information from censored observations. To facilitate imputation, a joint model is developed for the data, which includes a hierarchical change-point model for the time-dependent auxiliary variable and a time-dependent proportional hazards model. The MCMC methods are used to multiply impute event times for censored cases and then a standard analysis is conducted. The simulation study shows that the use of the MI method can lead to improved performance of estimators and the MCMC yielded less variable estimators which were closer to those produced by the fully observed method than were the estimates produced by the partially observed method. Aerts, *et al* (2002) studied a fully nonparametric and a semiparametric imputation method based on local resampling principles, which result in a consistent estimator under few or no parametric assumptions. Kernel methods for imputation of missing values were introduced by Titterington & Sedransk, (1989), who used kernel density estimation in combination with a nonparametric bootstrap for imputing values. For missing covariate data, smoothing methods have been applied by Wang, *et al* (1998) to estimate selection probabilities. Other semiparametric approaches, in the sense of not having to specify a fully parametric model, although not directly in a smoothing context, are constructed for drop-out models in Scharfstein, *et al* (1999). The authors went on to introduce two classes of local bootstrap methods; the fully nonparametric local resampling method which relaxes distributional assumptions and assumptions concerning regression functions and the local semiparametric method which assumes that the conditional distributions are locally normal but allows nonlinear conditional mean structures.

Zhang (2003) in a review paper entitled "Multiple Imputation: Theory and Method" discussed how to create proper imputations when data are missing. In the presence of missing data, three issues are of main concern; loss of efficiency, complications in data handling and potential serious bias due to the systematic differences between the observed and the missing data (Barnard & Meng, 1999). Barnard & Meng (1999) presents the three widely used multiple imputation methods; the propensity score method, the predictive model and the MCMC method.

Ghosh-Dastidar, *et al* (2003) presented a method called the Multiple Edit / Multiple imputation (MEMI), an extension of multiple imputation for handling the problems of nonresponse and response errors. The method replaces an observed data set containing missing values and errors with $m > 1$ simulated versions of the ideal data set that is complete and error-free. These ideal data sets are analysed separately, and the results are combined using the same rules as for multiple imputation. The resulting inferences simultaneously reflect uncertainty due to nonresponse and response errors. The MEMI may be an attractive alternative to deterministic or quasi-statistical edit and imputation procedures normally used.

Zio, *et al* (2004) presented a method to deal with the problem of the consistency of imputed values; preservation of statistical relationships between variables (statistical consistency) and preservation of logical constraints in data (logical consistency). This method is an extension to the work of Thibaudeau & Winkler (2002) bayesian networks for imputing missing values. Bayesian networks are useful for dealing with high dimensional statistical problems. Zio, *et al* (2004) allow a reduction in the complexity of the phenomenon under study by representing joint relationships between a set of variables through conditional relationships between subsets of these variables. Zio, *et al* (2004) however reported that an outstanding problem in the field of imputation is the preservation of joint relationships between variables.

Chen & Sun (2010) presented and investigated a multiple imputation approach when interval-censored data is generated under the additive hazards model. The authors claim that their approach is simple and easy to implement since it uses existing software packages for right-censored data analysis. Chen & Sun (2010) recommend rigorous tests in order to justify normality assumptions in making inferences about the covariates. Chen & Sun (2010) also suggest the development of a formal procedure for model comparison between the PHs and an additive hazards model.

2.6 Diagnostics

Farrington (2000) developed diagnostic tools for use with proportional hazard models for interval-censored life time data. Farrington (2000) proposed counterparts to the Cox-Snell (Cox & Snell, 1968), Lagakos (1980), (martingale), deviance (Therneau, *et al*, 1990), and Schoenfeld (Schoenfeld, 1982) residuals. Many of the properties of these residuals carry over to the interval-censored case; hence this work is an extension of the right-censored data counterparts. In particular, the interval-censored versions of the Lagakos (1980) and Schoenfeld (1982) residuals may be derived as components of suitable score statistics. The Lagakos (1980) residuals may be used to check regression relationships, while the Schoenfeld (1982) residuals assist to detect nonproportional hazards in semiparametric models.

Ren (2003) proposed the Cramér-von Mises type goodness of fit tests for interval censored data case 2 basing on a resampling method called the leveraged bootstrap (LB). The consistency of the method is also shown mathematically, with support from simulation studies. The main difference between the leverage bootstrap (LB) tests and the usual testing procedures is that the test statistics of the LB are obtained through resampling, in the process of which the leveraged bootstrap transfers censored data through some statistic T_n defined in the paper into some useful information from which to draw inference. Although the proposed tests can be applied to other types of censored data, they are intended mainly to fill the void for the interval-censored data. Simulation studies show that the proposed methods are efficient because EM algorithm is used only once in the procedure to compute the NPMLE. The proposed tests are computationally efficient, and are applicable to right-censored, doubly-censored and case k interval-censored data. Ren (2003) recommends the need for further research on the goodness of fit tests.

Chapter 3 ： Materials and methods

3.1 Introduction

Survival data is frequently presented as interval-censored. This form usually occurs in longitudinal studies, where the individuals are followed for a pre-fixed time period or visited periodically for a fixed number of times. The time, $T_i (i = 1, \ldots, n)$ until the occurrence of the event of interest for each individual is only known, whenever it occurs, to lie within the interval between visit times L_i and U_i, where L_i is the last time before occurrence and U_i is the first time after the occurrence of the event. It is only known that $L_i < T_i \leq U_i$. The survival time, T, is said to be interval-censored. If, however, the event occurs at the exact moment of a visit, though rare, then the time is said to be an exact survival time, and in this case $L_i = T_i = U_i$. In this study, in order to use right-censored methods to analyze interval-censored data, the times were first transformed into exact and right-censored times, using mid-point imputation. The method of mid-point imputation assumes that an event occurred at the mid-point of the interval in which it is recorded. After the transformation, the time to be used for analysis is given by $T_i = \frac{U_i + L_i}{2}$. It will be attempted to show that a very similar result to the interval-censored conventional methods can be obtained using the right-censored method after the mentioned transformation.

3.2 Materials

3.2.1 *The ACTG 181 data set description*

The AIDS Clinical Trials Network (ACTG), which is a funded by the National Institute of Health, was established in 1987. The group organizes and studies the prevention and treatment of HIV-1 infection.

In 1989, the ACTG initiated the ACTG 181 observational study. In this study, blood and urine were drawn from subjects at scheduled clinic visits. The visits were scheduled to allow monitoring for the time-to-shedding of the opportunistic infection cytomegalovirus (CMV) in an HIV-infected individual. The infection normally leads to blindness. The subjects provided blood and urine samples at clinic visits. Urine samples were supposed to be collected every 4 weeks while blood was to be drawn every 12 weeks. Many subjects had both samples taken every 4 weeks during the office visits and many visits were missed. The subject's CD4+ count, which acts as an indicator of HIV stage was dichotomised at baseline as to indicate whether or not a subject was in the late (less than 75 cells/µl of blood) or early (more than 75 cells/µl of blood) stage of the disease. For blood shedding, 7 subjects were left-censored, 23 interval-censored and 174 right-censored. For the urine shedding, 49 subjects were left-censored, 67 interval-censored and 88 right-censored. Subjects of the trial were drawn from patients, on a clinical trial (ACTG 081), who were randomised to receive one of the three treatment regimens to prevent *Pneumocystis carinii pneumonia* (Bozzette, *et al*, 1995). The marginal Cox method was used by Goggins & Finkelstein (2000) on the same problem and they also made comparisons against the

Finkelstein (1986) method. The data set is available online and details are given in Appendix B on page 77.

3.2.2 *Variable description*

The variables used in this study were as follows:

Obs subject observation number [lies between 1 and 234];

SEX gender [0 – female and 1 – male];

RACE ethnic groups [1 - white, 2 - black and 3 - other];

FIRSTCD4 earliest 181 CD4 count;

fcd4stat dichotomised earliest 181 CD4 count [0 – count less than 75 cells/µl and 1 – count greater than 75 cells/µl];

sheddind 181 CMV shedding indicator;

blposind 181 blood shedding indicator;

urposind 181 urine shedding indicator;

deathcen 081 death censor indicator;

BNEG duration (in 28 day units) from the earliest date of blood test to the last date of negative test $\left(\frac{number\ of\ days\ from\ BLOODEDT\ to\ BLDNEGDT}{28}\right)$ as a whole number;

BPOS duration (in 28 day units) from the earliest date of blood test to the first date of positive test $\left(\frac{number\ of\ days\ from\ BLOODEDT\ to\ BLDPOSDT}{28}\right)$as a whole number;

BOFF duration (in 28 day units) from the earliest date of blood test to the date the subject left the study $\left(\frac{\textit{number of days from BLOODEDT to BLDPOSDT}}{28}\right)$ as a whole number;

BSURVTM time spent on the study before the subject shed the virus in blood:

(i) left-censored observation BSURVTM = 0;

(ii) interval-censored time BSURVTM = $\frac{1}{2}$(BNEG + BPOS) ie $T_i = \frac{L_i + U_i}{2}$;

(iii) right-censored time BSURVTM = BNEG.

bcensind mid-point imputation censor indicator for shedding in blood [0 – right-censored times and 1 – event experienced];

UNEG duration (in 28 day units) from the earliest date of urine test to the last date of negative test $\left(\frac{\textit{number of days from URINEEDT to URNNEGDT}}{28}\right)$[as a whole number];

UPOS duration (in 28 day units) from the earliest date of urine test to the first date of positive test $\left(\frac{\textit{number of days from URINEEDT to URNPOSDT}}{28}\right)$ as a whole number;

UOFF duration (in 28 day units) from the earliest date of urine test to the date the subject left the study $\left(\frac{\textit{number of days from URINEEDT to OFFSTDT}}{28}\right)$ as a whole number;

USURVTM time spent on the study before the subject shed the virus in urine

(i) left-censored observation USURVTM = 0;

(ii) interval-censored time USURVTM = $\frac{1}{2}$(UNEG + UPOS) ie $T_i = \frac{L_i + U_i}{2}$;

(iii) right-censored time USURVTM = UNEG.

ucensind mid-point imputation censor indicator for shedding in urine [0 – right-censored times and 1 – event experienced].

3.3 Methods

The methods outlined below were used to analyse the mid-point imputed interval-censored data explained in ***Section 3.2.1.*** The data was analysed using the Kaplan-Meier estimator, the logrank test and a proportional hazards model was fitted.

3.3.1 *The Kaplan-Meier Product-Limit Estimator*

Before examining at parametric models for a set of data, it is often useful to explore the data by means of a nonparametric estimation procedure. The earliest and most commonly used method in survival data is the Kaplan-Meier product-limit estimator (Kaplan & Meier 1958).

If π_j is the probability of having an event at time t_j, conditional on not having an event until then, the likelihood function is given by;

$$L(\pi) = \prod_{j=1}^{k} \pi_j^{d_j} \left(1 - \pi_j\right)^{n_j - d_j} \tag{3.1}$$

where;

n_j is the conditional number having survived and still under observation, and hence still known to be at risk just prior to t_j, called the risk set,

d_j is the number having the event at t_j and

π_j is the hazard or intensity at t_j

k is the total number of death throughout the observation period.

This structure is a special application of the binomial distribution, with maximum estimates;

$$\hat{\pi}_j = {}^{d_j}\!/_{n_j} \tag{3.2}$$

Thus the product-limit estimator of the survivor function is just the product of the estimated probabilities of not having the event at all time points up to the one of interest;

$$\hat{S}(t) = \prod_{j|t_j<t} \left(1 - {}^{d_j}\!/_{n_j}\right) \tag{3.3}$$

This estimate of the survival distribution can be compared with known survival distributions using the logrank test to see if it follows any particular distribution. Should this be the case, the estimate parametric methods will be used to analyze the survival times.

3.3.2 *The Logrank Test*

The Kaplan-Meier survival curves are only used as descriptive and initial procedures of assessing the behaviour of the two groups. There is always need to formally test the hypothesis that the difference that may be observed through the Kaplan-Meier curves is not by chance, but is actually statistically significant. The most commonly used and formal test for comparison of survival times is the logrank test.

Suppose we have two groups of survival times, and we label the groups as Group I and II. We consider, separately, the death time in each of the two groups. Suppose that there are r distinct death times, then $t_{(1)} < t_{(2)} < t_{(3)} < ... < t_{(r)}$, across the two groups, and that at time t_j, d_{1j} individuals in Group I and d_{2j} individuals in Group II die, for $j = 1, 2, 3, ..., r$. Unless two or more individuals in a group have the same recorded death time, the values of d_{1j} and d_{2j} will either be zero or unity. Suppose further that there are n_{1j} individuals at risk of death in the first group just before time $t_{(j)}$, and that there are n_{2j} at risk in the second group. Then, at time $t_{(j)}$, there are $d_j = d_{1j} + d_{2j}$ deaths in total out of $n_j = n_{1j} + n_{2j}$ individuals at risk.

A way of assessing the null hypothesis of no difference between the two survival curves,

$$H_0 : S_1(t) = S_2(t) \tag{3.4}$$

is to consider the extent of the difference between the observed number of individuals in the two groups who die at each of the death times, and the numbers expected under the null hypothesis. This test can be presented as the conditional probability of observing d_{1j} deaths in Group I and d_{2j} deaths in Group II given that there are d_j deaths at time t_j. The conditional probability of observing d_{1j} deaths in Group I and d_{2j} deaths in Group II given that there are d_j deaths at tie t_j follows a hyper-geometric distribution defined in Breslow (1979) hence;

$$\Pr\left(d_{1j} \mid d_j, n_{1j}, n_{2j}\right) = \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}} \tag{3.5}$$

The mean, e_{1j} and variance v_{1j} of d_{1j} are given by the expressions;

$$e_{1j} = \frac{n_{1j}d_j}{n_j} \tag{3.6}$$

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^{\ 2}(n_j - 1)} \tag{3.7}$$

respectively.

The logrank test statistic U can then be defined as

$$U = \sum_1^r \left(d_{1j} - e_{1j}\right)$$

$$= \sum_1^r d_{1j} - \sum_1^r e_{1j}$$

$$U = O_1 - E_1 \tag{3.8}$$

where O_1 and E_1 are the observed event and expected event times for Group I and II, respectively.

Under the null hypothesis;

$$U \sim N\left(0, \sum_{j=1}^{r} v_{1j}\right) \tag{3.9}$$

Hence letting $V = \sum_{j=1}^{r} v_{1j}$ and standardizing yields;

$$\frac{U}{\sqrt{V}} \sim N(0,1) \tag{3.10}$$

Thus:

$$W = \frac{U^2}{V} \sim \chi_1^2 \tag{3.11}$$

The procedure was proposed by Mantel & Haenszel (1959), and is known as the Mantel-Haenszel procedure. The test based on this statistic has various names, including Mantel-Cox and Peto-Mantel-Haenszel, but is best known as the logrank test.

3.3.3 *The Cox Proportional Hazards Model*

The Cox regression model is usually given as;

$$\lambda_i(t) = \lambda_0(t) * \exp\left\{\beta_1 x_{j1} + ... + \beta_k x_{jk}\right\} \tag{3.12}$$

This equation says that the hazard for individual i at time t is the product of two factors;

- a baseline hazard function $\lambda_0(t)$ that is left unspecified, except that it cannot be negative
- a linear function of a set of k covariates, which is then exponentiated.

The baseline hazard function $\lambda_0(t)$ can be regarded as the hazard function for an individual whose covariates all have values of 0.

Taking logarithms of both sides will give the following;

$$\log\left\{\lambda_i(t)\right\} = \alpha(t) + \beta_1 x_{j1} + ... + \beta_k x_{jk} \tag{3.13}$$

where $\alpha(t) = \log\left\{\lambda_0(t)\right\}$. If we let $\alpha(t) = \alpha$, then we get the exponential model and if we let $\alpha(t) = \alpha t$, then we get the Gompertz model and finally if we let $\alpha(t) = \alpha \log t$, we have the Weibull model. Such choices when dealing with Cox regression are, however, unnecessary, thus $\alpha(t)$ can take any form.

The Cox model is called the proportional hazards model because the hazard for any individual is a fixed proportion of the hazard for any other individual (Allison, 1995). To visualise this relationship, if we take the ratio of the hazards for two individuals i and j, then we have:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp\left\{\beta_1\left(x_{i1}-x_{j1}\right)+\ldots+\beta_k\left(x_{ik}-x_{jk}\right)\right\} \tag{3.14}$$

Since the ratio of any two hazards is independent of time, and the baseline hazard, estimation of the baseline hazard is unnecessary. As a result, the ratio of the hazards is constant over time. Now letting X be an indicator variable, which takes the value zero if an individual is on the standard drug, and unity if the individual is on the new drug, say, and if x_i is the value of X for the ith individual in the study, $i = 1,2,3,\ldots,k$, then the hazard function for this individual can be written as;

$$\lambda_i(t) = \lambda_0(t)\exp\left\{\beta x_i\right\} \tag{3.15}$$

where $x_i = 1$ if the ith individual is on the new treatment and $x_i = 0$ otherwise.

3.3.3.1 Estimation procedures without tied survival times

Suppose that k of the survival times from n individuals are uncensored and distinct, and $n-k$ are right-censored. Let $t_1 < t_{(2)} < t_{(3)} < \ldots < t_{(k)}$ be the ordered k distinct failure times with corresponding covariate values $X_{(1)},\ldots,X_{(k)}$. Let $R\left(t_{(i)}\right)$ be the risk set at time $t_{(i)}$. $R\left(t_{(i)}\right)$ consists of all persons whose survival times are at least $t_{(i)}$. The partial likelihood function is, thus, given by:

$$L(\beta)=\prod_{i=1}^{k}\frac{\exp\left(\sum_{j=1}^{p}\beta_j X_{j(i)}\right)}{\sum_{l\in R\left(t_{(i)}\right)}\exp\left(\sum_{j=1}^{p}\beta_j X_{jl}\right)}=\prod_{i=1}^{k}\frac{\exp\left(\sum_{j=1}^{p}\beta_j^{'}X_{(i)}\right)}{\sum_{l\in R\left(t_{(i)}\right)}\exp\left(\sum_{j=1}^{p}\beta_j^{'}X_l\right)} \tag{3.16}$$

Letting $p=1$, the partial log likelihood is then given by:

$$l(\beta)=\sum_{i=1}^{k}\left\{\beta_j^{'}X_{(i)}-\log\left[\sum_{l\in R\left(t_{(i)}\right)}\exp\left(\beta_j^{'}X_l\right)\right]\right\} \tag{3.17}$$

The maximum partial likelihood estimator (MPLE), $\hat{\beta}$, of β can be obtained by solving the equation;

$$\frac{\delta\left[l(\beta)\right]}{\delta\beta}=0 \tag{3.18}$$

by applying the Newton-Raphson iterated procedure. The second partial derivatives of $l(\beta)$ with respect to β_u and β_v, $u,v=1,...,p$ using the Newton-Raphson iterative procedure are;

$$I_{uv}(\beta)=\frac{\delta^2\left[l(\beta)\right]}{\delta\beta_u\delta\beta_v}=-\sum_{i=1}^{k}C_{(uvi)}(\beta) \tag{3.19}$$

where

$$C_{(uvi)}(\beta)=\frac{\sum_{l\in R\left(t_{(i)}\right)}X_{ul}X_{vl}\exp\left(\sum_{j=1}^{p}\beta_j X_{jl}\right)}{\sum_{l\in R\left(t_{(i)}\right)}\exp\left(\sum_{j=1}^{p}\beta_j X_{jl}\right)}$$

$$-\frac{\left(\sum_{l\in R\left(t_{(i)}\right)} X_{ul}\exp\left(\sum_{j=1}^{p}\beta_j X_{jl}\right)\right)\left(\sum_{l\in R\left(t_{(i)}\right)} X_{vl}\exp\left(\sum_{j=1}^{p}\beta_j X_{jl}\right)\right)}{\left(\sum_{l\in R\left(t_{(i)}\right)}\exp\left(\sum_{j=1}^{p}\beta_j X_{jl}\right)\right)^2}$$

(3.20)

The covariance matrix of the MPLE $\hat{\beta}$ is defined by:

$$\hat{V}(\hat{\beta}) = Cov(\hat{\beta}) = \left[-\frac{\delta^2\left[l\left(\hat{\beta}\right)\right]}{\delta\beta\delta\beta^{'}}\right]^{-1} \tag{3.21}$$

where the term $\left[-\frac{\delta^2\left[l\left(\hat{\beta}\right)\right]}{\delta\beta\delta\beta^{'}}\right]$ is called the *observed information matrix.*

Let the (i, j) element of $\hat{V}(\hat{\beta})$ be $v_{i,j}$; then the marginal $100(1-\alpha)\%$ confidence interval for β_i is given by:

$$\left(\beta_i \pm Z_{\alpha/2}\sqrt{\hat{V}(\hat{\beta})}\right) \tag{3.22}$$

3.3.3.2 Estimation procedures with tied survival times

Suppose that among the n observed survival times there are k distinct uncensored times, let $t_{(1)} < t_{(2)} < t_{(3)} < ... < t_{(k)}$. Let $m_{(i)}$ denote the number of people who fail at $t_{(i)}$ or the multiplicity of$t_{(i)}$; $m_{(i)} > 1$ if there is more than one observation with value$t_{(i)}$; $m_{(i)} = 1$ if there is only one observation with value $t_{(i)}$. Let $R\left(t_{(i)}\right)$ denote the set of people at risk at time $t_{(i)}$ [thus, $R\left(t_{(i)}\right)$consists of those whose survival times are at least $t_{(i)}$ and r_i be the number of such persons.

From every $R\left(t_{(i)}\right)$, we can randomly select $m_{(i)}$ subjects. If we denote each of these $m_{(i)}$ by U_j, then there are $\begin{pmatrix} r_i \\ m_{(i)} \end{pmatrix} = \frac{r_i!}{m_{(i)}!(r_i - m_{(i)})!}$ possible U_js. Let U_i denote the set that contains all the U_js. Focusing on the tied observations; Let $X_k = \left(X_{1k},...,X_{pk}\right)^{'}$denote the covariates of the kth individual, $Z_{U_{(j)}} = \sum_{k \in U_{(j)}} X_k = \left(Z_{1U_{(j)}},...,Z_{pU_{(j)}}\right)^{'}$, where $Z_{lU_{(j)}}$ is the sum of the lth covariate of the $m_{(i)}$ subjects who are in U_j. Let U_i* denote the set of $m_{(i)}$ subjects who failed at time $t_{(i)}$, and $Z_{U_{(j)}*} = \sum_{k \in U_{(j)}*} X_k = \left(Z*_{1U_{(j)}*},...,Z_{pU_{(j)}*}\right)^{'}$, where $Z*_{lU*_{(j)}}$ is the sum of the lth covariate of the $m_{(i)}$ subjects who are in $U*_{(i)}$ (failed at time $t_{(i)}$).

For the continuous time scale, to approximate the exact partial likelihood function, we can use the function provided by Breslow (1974);

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp\left(Z^{'}_{U_{(i)}*}\beta\right)}{\left[\sum_{l \in R\left(t_{(i)}\right)} \exp\left(X^{'}_{l}\beta\right)\right]^{m_{(i)}}} \tag{3.23}$$

and an alternative approximation was provided by Efron (1977);

$$L_E(\beta)=\prod_{i=1}^{k}\frac{\exp\left(Z'_{U_{(i)}*}\beta\right)}{\prod_{i=1}^{m_{(i)}}\left\{\sum_{l\in R\left(t_{(i)}\right)}\exp\left(X'_l\beta\right)-\left[\frac{j-1}{m_{(i)}}\right]\sum_{l\in U_{(i)}*\left(t_{(i)}\right)}\exp\left(X'_l\beta\right)\right\}} \tag{3.24}$$

If survival times are observed at discrete times, the tied observations are true ties then the partial likelihood function at the discrete time scale is given by:

$$L_d(\beta)=\prod_{i=1}^{k}\frac{\exp\left(Z'_{U_{(i)}*}\beta\right)}{\left\{\sum_{U_{(j)}}\exp\left(Z'_{U_{(i)}}\beta\right)\right\}} \tag{3.25}$$

3.3.4 *Model checking*

3.3.4.1 The Cox-Snell residuals

The most widely used residual in analysis of survival data is the Cox-Snell residual, because it is a particular example of the general definition of residuals given in Cox & Snell (1968), (Collett, 1994).

The Cox-Snell residual for the ith individual, $i=1,...,n$ is given by;

$$r_{C_i}=\exp\left(X'_i\hat{\beta}\right)\hat{\Lambda}_0(t_i) \tag{3.26}$$

where $\hat{\Lambda}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time of the individual.

When the correct model is fitted, the r_{C_i} will follow a unit exponential distribution and both the mean and variance of the ith residual will both be unity. The residuals cannot be negative, hence, cannot be symmetric about zero. In addition, a point to note is that if the largest survival time is uncensored, the estimated value of the survival function beyond that time is zero, and r_{C_i} is undefined for that observation.

3.3.4.2 The Modified Cox-Snell residuals

Suppose that the ith survival time is a censored observation, and let $t_i *$, and let t_i be the actual, but unknown, survival time, so that $t_i > t_i *$. The Cox-Snell residual for this individual, evaluated at the censored survival time, is given by;

$$r_{C_i} = \hat{\Lambda}_i(t_i*) = \log \hat{S}_i(t_i*), \qquad (3.27)$$

where $\hat{\Lambda}_i(t_i*)$ and $\hat{S}_i(t_i*)$ are the estimated cumulative hazard and survival functions, respectively for the ith individual at the censored survival time.

The modified version of the Cox-Snell residual is given by;

$$r_{C_i}{}' = \begin{cases} r_{C_i} & for \quad observed \quad times \\ r_{C_i} + \Delta & for \quad censored \quad times \end{cases} \tag{3.28}$$

where r_{C_i} is the Cox-Snell residual for the ith observation. Since r_{C_i} has a unit exponential distribution, the excess residual, Δ will also have the same distribution with the expected value of Δ being unity, and this leads to the modified Cox-Snell residual as;

$$r_{C_i}{}' = \begin{cases} r_{C_i} & for \quad observed \quad times \\ r_{C_i} + 1 & for \quad censored \quad times \end{cases} \tag{3.29}$$

Crowley and Hu (1977) suggested that the use of the mean of the excess residual tends to inflate the residual and suggested the use of the median, and came up with the modified Cox-Snell residual given by;

$$r_{C_i}{}'' = \begin{cases} r_{C_i} & for \quad observed \quad times \\ r_{C_i} + \log 2 & for \quad censored \quad times \end{cases} \tag{3.30}$$

3.3.4.3 The Martingale residuals

The Cox-Snell modified residuals $r_{C_i}{}'$ have a mean of unity for uncensored observations and can further be refined so that transformed residuals have a mean of zero when an observation is uncensored. The Cox-Snell residuals can thus be modified to get the martingale residuals (derivable from martingale methods);

$$r_{M_i} = \delta_i - r_{C_i} \tag{3.31}$$

The martingale residuals take values between $-\infty$ and unity with residuals for censored observations being negative. The martingale residuals are not symmetrically distributed about zero.

3.3.4.4 The deviance residuals

The martingale residuals given above lack symmetry about zero hence are skewed. The skewness, even in the presence of a correct model fitted, makes plots based on the residuals difficult to interpret. The deviance residuals, introduced by Schoenfeld (1982) are more symmetric about zero.

The deviance residual is defined as;

$$r_{D_i} = sign(r_{M_i})\sqrt{-2\left[r_{M_i} + \delta_i \log(\delta_i - r_{M_i}\right]} \tag{3.32}$$

where r_{M_i} is the martingale residual for the ith individual, and the function sign() is the sign function. When a good model has been fitted, the deviance residuals can be expected to be symmetrically distributed about zero, but do not necessarily sum to zero. In this study, deviance residuals were used to check the goodness-of-fit of the model.

Chapter 4 : Results and Analysis

4.1 Introduction

The data set ACTG 181 was analysed using SAS 9.2. The ACTG 181 subjects were categorized at entry into the study as having either a high CD4+ cell count (more than 75cells/µl), or low CD4+ cell count (less than 75cells/µl). Subjects with a high cell count were considered to be in their early stage of HIV while those with a low cell count were in their late stage. The proc lifetest option of the SAS 9.2 software was used to plot the Kaplan-Meier curves and calculate the logrank test to compare survival times between subjects in the two disease stages (early and late). There was a significant difference in time-to-shedding, in both urine and in blood, between subjects in the early stage as compared to those in the late stage at entry. Since the Kaplan-Meier curves did not cross each other, the assumption of proportional hazards can be made and hence the Cox proportional hazards model was fitted to the data to assess the hazard ratio between the two disease stages at entry into the study.

4.2 Results

4.2.1 *Frequencies*

The ACTG 181data set consisted of 232 HIV-1 infected patients on treatment. For analysis, patients had to have any early date of CMV observation in either blood or urine. However, 28 subjects in the study did not have any early CMV observation dates so the subjects were excluded. Only 204 of the subjects were considered for analysis in this study. 96% of the subjects were males. Of the 204 subjects, 178 were white while only 11 were black and the rest belonged to other ethnic groups. The study was set to assess the hazard of the disease stage as early or late at entry into the study. The distribution of the dichotomized disease stage (variable fcd4stat) was almost balanced (Table 4.1, page 51).

Table 4.1 : Baseline characteristics of the 204 HIV-infected subjects by disease stage

Characteristic	fcd4stat=0 (N=111)	fcd4stat=1 (N=93)	Total (N=204)
*Disease stage	Late	Early	
Race - No. of subjects			
white	94(84.7%)	84(90.3%)	178(87.3%)
black	14(12.6%)	9(9.7%)	23(11.3%)
other	3(2.7%)	0(0.0%)	3(1.4%)
Gender - No. of subjects			
male	106(95.5%)	87(93.5%)	193(94.6%)
female	5(4.5%)	6(6.5%)	11(5.4%)
CD4+ count (cells/ml)			
mean	25.7	206.2	
median	20	172	

*Subjects in the early HIV disease stage had higher CD4+ counts than later.

The distribution of baseline characteristics was compared across the dichotomized CD4+ cell count at entry, coded as fcd4stat=0 for a CD4+ cell count less than 75 cells/µl and fcd4stat=1 for otherwise (Table 4.1, page 51). However, the mean and median CD4+ cell count was higher for subjects in the early stage of the disease.

Shedding was experienced by 78 (38.2%) subjects. 10 subjects shed in blood alone, 56 subjects shed in urine alone while 12 subjects experienced shedding in both blood and urine. All the 22 subjects who eventually shed in blood were white males (Table 4.2, page 53). More than 60% of the subjects who experienced shedding were in their late stage of the disease (68.2% for blood and 60.3% for urine).

Table 4.2 : Distribution of CMV shedding in blood and urine by baseline variables

Characteristic	shedding in blood alone N=10	shedding in urine alone N=56	shedding in both N=12	No shedding N=126
Race - No. of subjects				
white	10(100.0%)	47(84.0%)	12(100.0%)	109(86.5%)
black	0(0.0%)	9(16.0%)	0	14(11.1%)
other	0(0.0%)	0(0.0%)	0	3(2.4%)
Gender - No. of subjects				
male	10(100.0%)	55(98.2%)	12	116(92.1%)
female	0(0.0%)	1(1.8%)	0	10(7.9%)
CD4+ count (cells/µl) at entry				
<75	15(68.2%)	41(60.3%)		
>75	7(31.8%)	27(39.7%)		
CD4+ count (cells/µl)				
mean	59.9	94.9		
median	26.5	49		

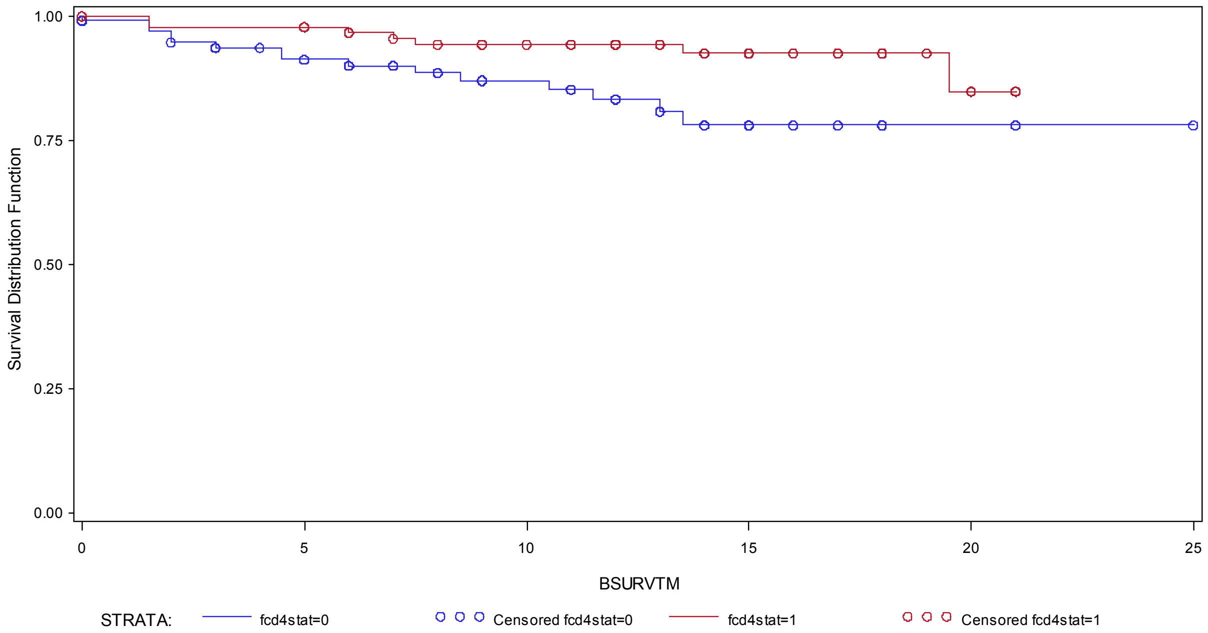
For analysis of subject by shedding/non-shedding of CMV by disease stage, see Appendix E (Page 81) and Appendix F (Page 82).

4.3 Analyses

4.3.1 *Kaplan-Meier curves*

The SAS 9.2 proc lifetest procedure was used to make a comparative analysis on the survival times between subjects who, at entry into the study, were at the late stage of HIV (fcd4stat=0) against those in the early stage (fcd4stat=1).

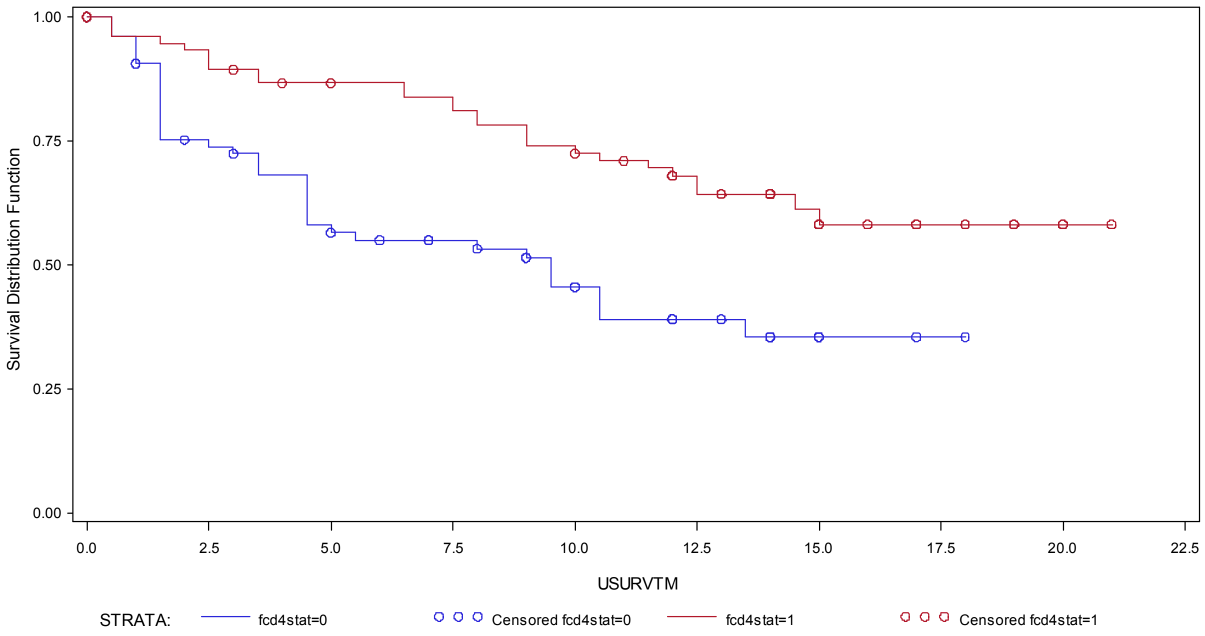
Time-to-shedding of CMV in blood produced the Kaplan-Meier curves in Figure 4.1 on page 55. From the curves, the graph for subjects in their late disease stage at entry (in blue) was consistently lower than that for subjects in their early stage at entry (in red). There was a gradual increase in the difference from the onset of the study up to around the 18th month, when the difference suddenly dropped. For subjects in their late disease stage at entry, the first subject to experience shedding in blood did so at entry and the last (15th) subject to experience shedding did so after 13.5 months. The median value for the subjects in their late disease stage experienced shedding in the 7th month. As for the subjects in their early disease stage at entry, the first subject to shed in blood did so after 1.5 months in the trial and the last (7th) shed after 19.5 months in the study. In both cases, the number of subjects who experienced shedding in blood was less than 25%. However, of the subjects in their early disease stage who experienced shedding in blood the median value was the 4th subject after about 7 months in the study. The logrank test for homogeneity was calculated to test the hypothesis of equal time-to-shedding between the two stages of the disease at entry. The result showed statistically significant evidence against equality [χ^2 = 5.2130 (1df) (P=0.0224)].



Key: fcd4stat=0 - early disease stage, fcd4stat=1 - late disease stage.

Figure 4.1 : Kaplan-Meier curves for time-to-shedding in blood against disease stage at entry into the ACTG 181 study

The Kaplan-Meier curves in Figure 4.2, page 56, resulted from the time-to-shedding of CMV in urine against the HIV stage at entry into the study. The two curves, did not cross each other with the one for the subjects in the late disease stage at entry (in blue) consistently below that of those who entered the study in their early stage (in red). The plot for the subjects in the early disease stage at entry had a gradual slope to the last subject. The first subject in the early disease stage at entry into the study experienced shedding in urine at 0.5 months into the study and the last (27th) subject experienced the event after 15 months into the study. The median could not be calculated because less than 50% of the subjects experienced the event. The 25th percentile was 9.0 time units [95% CI= (6.5, 12.5)]. As for the curve for the subjects in the late disease stage at entry, the slope was a steep drop up to about 5.0 months into the study, then the slope stabilized thereafter to the last subject. The first subject experienced shedding in urine at time 0.5 months and the last (41st) subject's time was 13.5 months in the trial. The survival times gave a median of 9.5 time units [95% CI = (4.5, 13.5)] and a 25th percentile value of 2.5 time units [95% CI = (1.5, 4.5)]. The 25th percentile survival time for subjects in the late disease stage at entry was lower than that for subjects in the early disease stage at entry in the study. The logrank test for homogeneity, gave a significant result [χ^2= 12.0827 (1df) (p-value = 0.0005)] and confirmed that subjects in the late disease stage were at a higher risk of shedding the CMV virus.



Key: fcd4stat=0 - early disease stage, fcd4stat=1 - late disease stage.

Figure 4.2 : Kaplan-Meier curves for time-to-shedding in urine against disease stage at entry into the ACTG 181 study

The difference in the times-to-shedding was more pronounced in time-to-urine shedding as compared to the time-to-blood shedding of the CMV. In both sets of curves, the curves for subjects in the late disease stage at entry were lower than that for subjects in the early disease stage at entry in the study.

4.3.2 *Cox proportional hazards model*

The SAS 9.2 phreg procedure was used to fit the model with three candidate explanatory covariates, SEX, RACE and fcd4stat. The stepwise selection criterion gave a significant parameter for fcd4stat in times to CMV shedding for blood as well as urine.

Table 4.3, page 58, gives a summary of the Stepwise Selection for the time-to-shedding of CMV in blood. The first step entered the variable fcd4stat, in Step 2 the variable RACE was entered then removed in Step 3. The process terminated when the variable RACE was removed with χ^2= 1.01044 (P=0.00288), giving a hazard ratio of 2.75. An analysis of the excluded variables gave a non-significant value χ^2= 0.5398 (P=0.4625) for SEX and χ^2= 2.6584 (P=0.1030) for RACE.

Table 4.3 : Stepwise Selection steps for the proportional hazards model for CMV-shedding in blood

Step (Covariance)	Criterion	Without Covariate	With Covariates
1. fcd4stat	AIC	214.999	211.803
2. RACE + fcd4stat	AIC	214.999	**208.373**
3. fcd4stat	AIC	214.999	211.803

The Stepwise Selection process for time-to-CMV shedding in urine entered fcd4stat in Step 1, SEX in Step 2 and removed SEX in Step 3 (See Table 4.4, page 59). Since the first lowest value of AIC is recorded in Step 1, fcd4stat is the only significant candidate variable for the model. The selection terminated in Step 3 with χ^2=0.83446 (P=0.0009), which gives a hazard ratio of 2.30. An analysis of the excluded variables gave a non-significant value χ^2 = 1.4551 (P=0.2277) for SEX and χ^2 = 0.0004 (P=0.9851) for RACE.

Table 4.4 : Stepwise Selection steps for the proportional hazards model for urine

Step (Covariance)	Criterion	Without Covariate	With Covariates
1. fcd4stat	AIC	627.806	**618.454**
2. fcd4stat + SEX	AIC	627.806	618.509
3. fcd4stat	AIC	627.806	618.454

It can be understood, from the two tables that time-to-CMV shedding was significantly predictable using disease stage at entry into the trial. Both results showed that subjects who came into the study in the late disease stage (less than 75 cells/μl) were more than twice as likely to experience shedding either in blood or in urine as the subjects coming into the study still in their early disease stage.

4.3.3 *Model Checking*

The adequacy of the model was checked using the deviance residuals.

For time-to-CMV shedding in blood, the residuals were generally centred at zero but there was one outlier recorded for subjects in the late disease stage at entry (Figure 4.3, page 60). Hence the models fitted for the CMV shedding data were the best models.

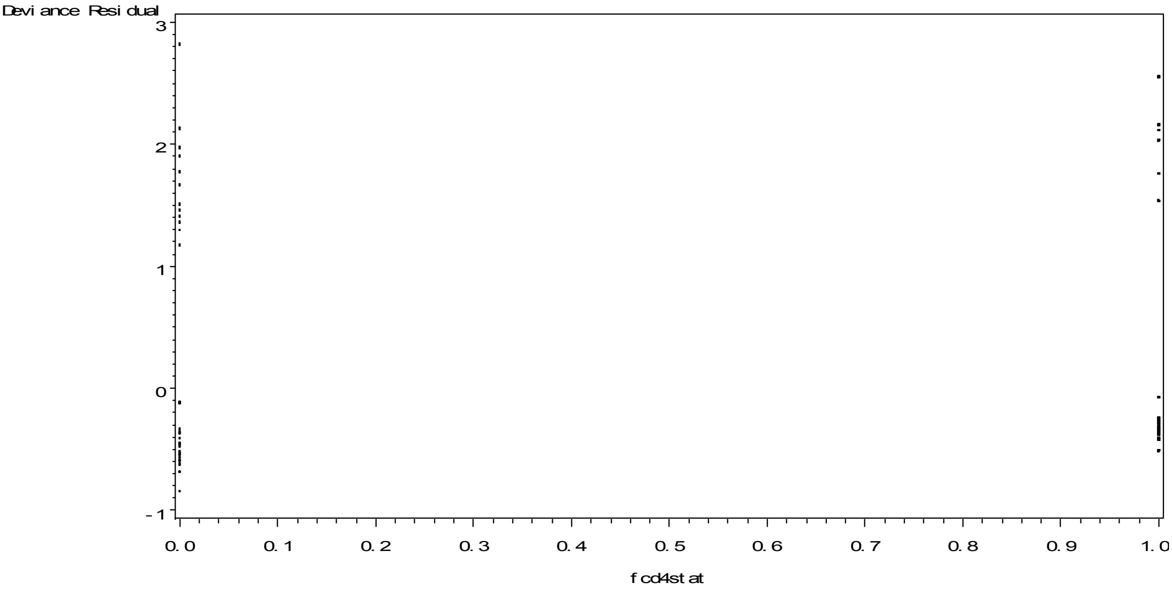


Figure 4.3 : Deviance residuals plots for disease stage (blood)

For time-to-CMV shedding in urine, there were no outliers and the plots were generally symmetrical about zero (Figure 4.4, page 61).

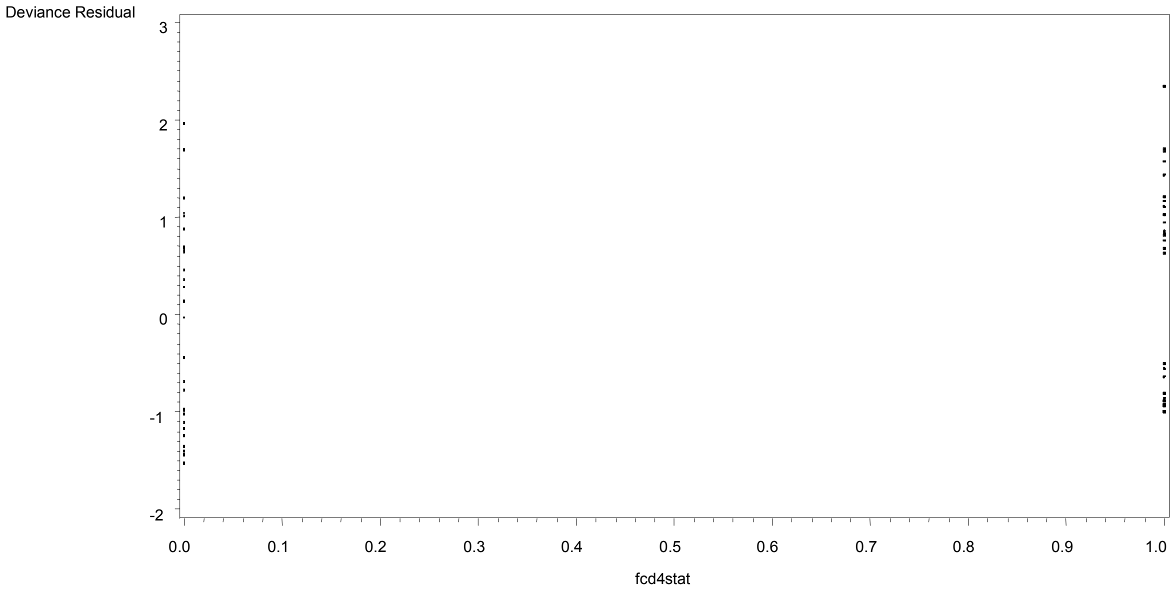


Figure 4.4 : Deviance residuals plots by disease stage (urine)

4.3.4 *Comparison of the mid-point imputation model with results from other methods*

The method of mid-point imputation was also used to model the ACTG 181 data set which was previously analysed using the univariate method of Finkelstein (1986) as well as the end-point imputation method of Cox (Cox, 1972). The results are shown in Table 4.5, page 62. The figures show that the results obtained using mid-point imputation are comparable with both the univariate Finkelstein (1986) and end-point imputation. The advantage of mid-point imputation is that mid-point imputation enables the researcher to use right-censored methods for interval-censored data. The method of mid-point imputation proposed in this study, however, tends to result in a slightly larger error in comparison to the other two, resulting in a wider confidence interval.

Table 4.5 : Comparison of results with other methods

Method	$\hat{\beta}_{blood}$ $\hat{\beta}_{urine}$	*s.e. $(\hat{\beta}_{blood})$ s.e. $(\hat{\beta}_{urine})$	Hazard Ratio	95% C. I. (Hazard Ratio)
Cox (1972)	0.90	0.41	2.46	(1.10, 5.49)
(end-point)	0.67	0.19	2.41	(1.35, 2.84)
Finkelstein (1986)	1.09	0.30	2.97	(1.65, 1.68)
(univariate)	0.88	0.20	2.41	(1.63, 3.57)
Cox (Mid-point)	**1.01**	**0.46**	**2.75**	**(1.11, 6.76)**
	0.83	**0.25**	**2.30**	**(1.40, 3.74)**

NB: *s.e. – standard error

Chapter 5 : Conclusion

5.1 Introduction

The aim of this study was to investigate the comparability of the right-censored methods, which included fitting the Cox proportional hazards model, on interval-censored data against the interval-censored methods. This comparison involved transforming interval-censored and left-censored observations into right-censored data using a method called mid-point imputation. The method asigns that the occurrence of an event within an interval to the mid-point of the particular interval. Left-censored observations were treated as right-censored at entry into the study and interval-censored observations were given the mid-points of their intervals as the imputed values. An observational study, the ACTG 181 was used to illustrate the method and the conclusions drawn from this study are presented below.

5.2 Discussion

From the results presented in Table 4.5, page 62, it was seen that the hazard values calculated using the mid-point imputation were lower than those values resulted from the univariate method but slightly higher than the values from the end-point method. This pattern applied to times-to-shedding in both blood and urine. It was seen that when censoring was severe (89.2%), in time-to-shedding of CMV in blood, the mid-point imputation method gave a larger standard error, requiring a wider confidence interval. In the time-to-shedding of CMV in urine, censoring was moderate (66.7%). Even though the standard error was still large, the difference was notably reduced. However, a confirmatory test showed that there was no evidence for a significant difference between the regression coefficients in both cases (Z=0.1053, P=0.9161 for time-to-shedding in blood and Z=0.1111, P=0.9115 for the time-to-shedding in urine).

Taking the Finkelstein (1986) model to be the yardstick, the mid-point imputation method yields better results than the endpoint method (Cox (1972) in Table 4.5, page 62) and the results are closer to those from the interval-censored methods especially when censoring is not very marked.

5.3 Areas of future research

Even though the method of mid-point imputation may seem very attractive, there is need for a number of issues to be investigated for one to comfortably use it. Using mid-point imputation to handle interval-censored data makes the assumption that the event time is known, hence with such an assumption and a deluge of many others, a number of issues need to be further examined.

From this study we note that the severity of censoring may have an effect on the accuracy of the regression model. Handling interval-censored data depends on making accurate judgements on the unknown lengths of intervals as well as observing the variations of the lengths of the intervals. Odell, *et al* (1992) point out that mid-point imputation could produce biased survival estimators especially at early points in the study. Interval-censored data may have left-truncated observations. Comparison of the performance of right-censored methods can also be assessed with varying levels of censoring, as well as the timing of censoring; early or late. There may be need to study methods to effectively handle such situations in conjunction with mid-point imputation. Simulation studies can usefully address these issues as part of future research on mid-point imputation.

5.4 Conclusion

In this study, right-censored methods were used to model interval-censored data. Mid-point imputation was used to convert the interval-censored observation into right-censored. A comparison of the results with the conventional interval-censored method showed that there was no significant difference in the results. However, it should be noted that the method generally underestimates the hazard ratio. Further research could be done to determine the conditions under which the methods could give the same results. The model fitted to the data showed that dichotomised CD4+ cell count level is predictive of the shedding of CMV.

Bibliography

Aerts, M., Claeskens, G., Hens, N. and Molenberghs, G. 2002. Local Multiple Imputation. *Biometrika* **89:** 375-388.

Allison, P. D. 1995. *Survival Analysis Using the SAS System: A Practical Guide.* SAS Press.

Altman, D. G. 1991. *PRACTICAL STATISTICS FOR MEDICAL RESEARCH.* Chapman & Hall. USA.

Barnard, J., and Meng, X. L. 1999. Small-sample degrees of freedom with multiple imputation. *Biometrika* **86:** 948-955.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. 1999. Local EM estimation of the hazard function for interval-censored data. *Biometrics* **55:** 238-245.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. 2002. A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine* **21:** 263-275.

Bozzette, S. A., Finkelstein, D. M., Spector, S. A., Frame, P., Powderly, W. G., He, W., Phillips, R. N., Craven, D., van der Horst, C. and Feinberg, J. 1995. A randomized trial of three antipneumocystis agents in patients with advanced Human Immunodeficiency Virus Infection. *The New England Journal of Medicine* **332:** 693-699.

Braun, J., Duchesne, T. and Stafford, J. E. 2005. Local Likelihood Density Estimation for Interval Censored Data. *Statistical Society of Canada* **33:** 39-60.

Breslow, N. 1974. Covariance Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review* **43:** 43-54.

Breslow, N. 1979. Statistical methods for censored survival data. *Environmental Health Perspectives* **32:** 181-192.

Breslow, N. and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistics Association* **88:** 9-25.

Cai, T. and Betensky, R. A. 2003. Hazard Regression for Interval-Censored Data with Penalized Spline. *Biometrics* **59:** 570-579.

Chen, L. and Sun, J. 2010. A multiple imputation approach to the analysis of interval-censored failure time data with additive hazards model. *Computational Statistics and Data Analysis* **54:** 1109-1116.

Clark, T. G., Bradburn M. J., Love, S. B. and Altman, D. G. 2003. Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer* **89:** 232 – 238.

Collett, D. 1994. Modelling Survival Data in Medical Research. *Chapman & Hall/CRC.* London.

Cox D. R. and Snell, E. J. 1968. A general definite of residuals. *Journal of the Royal Statistical Society, Series B* **30:** 248-275.

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

Cox, D. R. and Oakes, D. 1984. *Analysis of Survival Data.* Chapman & Hall/CRC. London.

Crowley, J. and Hu, M. 1977. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* **72:** 27-36.

DeGruttola, V. and Lagakos, S. W. 1989. Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45:** 1-11.

Dempster, A., Laird, N. and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B:* 1-38.

Deng, D. and Fang, H-B. 2009. Asymptotics for non-parametric likelihood estimation with doubly censored multivariate failure times. *Journal of Multivariate Analysis* **100:** 1802-1815.

Efron, B. 1967. The two sample problem with censored data. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, June 21-July 18, 1965; December 27, 1965-January7, 1966, Volume 4; Biology an Problems of Health* (LeCam, L. M. & Neyman, J. Eds.), University of California Press, Berkeley, 831-853.

Efron, B. 1977. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of American Statistical Association* **72:** 557-565.

Fang, H-B. and Sun, J. 2001. Consistency of nonparametric maximum likelihood estimation of a distribution function base on doubly interval-censored failure time data. *Statistics & Probability Letter* **55:** 311-318.

Farewell, V. T. 1986. Mixture models in survival analysis: Are they worth the risk? *The Canadian Journal of Statistics* **14:** 257-262.

Farrington, C. P. 2000. Residuals for Proportional Hazards Models with Interval-Censored Survival Data. *Biometrics* **56:** 473-482.

Faucett, C. L., Schenker, N. and Taylor, M. G. 2002. Survival Analysis Using Auxiliary Variables Via Multiple Imputation, with Application to AIDS Clinical Trial Data. *Biometrics* **58:** 37-47.

Finkelstein, D. M. 1986. A proportional hazards model for interval-censored failure time data. *Biometrics* **42:** 845-854.

Finkelstein, D. M., Goggins, W. B. and Schoenfeld, D. A. 2002. Analysis of failure time data with dependent interval censoring. *Biometrics* **58:** 298-304.

Gentleman, R. and Vandal, A. C. 2001. Computational Algorithms for Censored-Data Problems Using Intersection Graphs. *Journal of Computational and Graphical Statistics* **10:** 403- 421.

Ghosh-Dastidar, B. and Joseph L. Schafer, J. L. 2003. Multiple Edit/Multiple Imputation for Multivariate Continuous Data. *Journal of the American Statistical Association* **98:** 807- 817.

Goggins, W. B. and Finkelstein, D. M. 2000. A Proportional Hazards Model for Multivariate Interval-Censored Failure Time Data. *Biometrics* **56:** 940-943.

Gomez, G. and Calle, M. L. 1999. Nonparametric estimation with doubly censored data. *Journal of Applied Statistics* **26:** 45-58.

Gomez, G. and Lagakos, S. W. 1994. Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics* **50:** 204-212.

Grooeneboom, P. and Wellner, J. A. 1992. Information Bounds and Nonparametric Maximum Likelihood Estimation. *DMV Seminar, Brand 19, Birkhauser, New York.*

Gumbel, E. J. 1960. Bivariate exponential distributions. *Journal of the American Statistical Association* **55:** 698-707.

Hoff, R. 1994. Preparation for HIV Vaccine Trails: Moving From Baseline Studies to Efficacy Trials. *AIDS Research and Human Retroviruses* **10, supp. 2:** S191-S193.

Hudgens, M. G., Satten, G. A. and Longini, I. M. Jr. 2001. Nonparametric Maximum Likelihood Estimation for Competing Risks Survival Data Subject to Interval Censoring and Truncation. *Biometrics* **57:** 74-80.

Kalbfleisch, J. D. and Prentice, R. L. 1980. *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

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

Kim, D. K., DeGruttola, V. G. and Lagakos, S. W. 1993. analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49:** 13-22.

Komárek, A. and Lesaffre, E. 2006. Bayesian semi-parametric accelerated failure time model for paired doubly interval-censored data. *Statistical Modelling* **6:** 3-22.

Komárek, A. and Lesaffre, E. 2008. Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association* **103:** 523-533.

Kooperberg, C., Stone, C. J., and Truong, Y. K. 1995. Hazard regression. *Journal of the American Statistical Association* **90:** 78-84.

Kroner, B. L., Rosenberg, P. S., Aledort, L. M., Alvord, W. G. and Goedert, J. J. 1994. HIV-1 Infection Incidence Among Persons with Hemophilia in the United States and Western Europe, 1978-1990. *Journal of Acquired Immune Deficiency Syndrome* **7:** 279-286.

Kuk, A. Y. C. and Chen, C. 1992. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79:** 531-541.

Lagakos, S. W. 1979. General Right Censoring and Its Impact on the Analysis of Survival Data. *Biometrics* **35:** 139-156.

Lagakos, S. W. 1980. The graphical evaluation of explanatory variables in proportional hazards regression. *Biometrics* **68:** 93-98.

Lee, E. T. and Wang, J. W. 2003. *Statistical Methods for Survival Data Analysis* (3rd Edition). John Wiley and Sons, Inc., Hoboken, New Jersey.

Lesaffre, E., Komárek, A. and Declerk, D. 2005. An overview of methods for interval-censored data with an emphasis on applications in dentistry. *Statistical Methods in Medical Research* **14:** 539–552.

Li, L., Watkins, T. and Yu, Q. 1997. An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics* **24:** 531-542.

Mantel, N. and Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22:** 719-748.

Odell, P. M., Anderson, K. M. and D'Agostino, R. B. 1992. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* **48:** 951-959.

Pan, W. 2000. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56:** 199-203.

Peng, Y. and Dear, K. B. G. 2000. A Nonparametric Mixture Model for Cure Rate Estimation. *Biometric*s **56:** 237-243.

Peto, R. 1973. Experimental survival curves for interval-censored data. *Applied Statistics* **22:** 86-91.

Peto, R. and Peto, J. 1972. Assymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society A* **135:** 185-198.

Prentice, R. L. and Gloeckler, L. A. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34:** 57-67.

Ren, J. 2003. Goodness of Fit Tests with Interval Censored Data. *Scandinavian Journal of Statistics* **30:** 211-226.

Rosenberg, P. S., Goedert, J. J. and Biggar, R. J. 1994. Effect of age at seroconversion on the natural AIDS incubation distribution. *AIDS* **8:** 803-810.

Rubin, D. B. 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72:** 538-543.

Rubin, D. B. 1978. Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association* **3:** 20-34.

Satten, G. A. 1996. Rank-based Inference in the Proportional Hazards Model for Interval Censored Data. *Biometrika*, **83:** 355-370.

Satten, G. A., Datta, S. and Williamson, J. M. 1998. Inference Based on Imputed Failure Times for the Proportional Hazards Model with Interval-Censored Data. *Journal of the American Statistical Association* **441:** 318-327.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse model (With discussion). *Journal of the American Statistical Association* **94:** 1096-1146.

Schoenfeld, D. 1982. Partial residuals for the proportional hazards regression model. *Biometrika* **69:** 239-241.

Taylor, J. M. G. 1995. Semi-parametric estimation in failure time mixture models. *Biometrics* **51:** 899-907.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. 1990. Martingale-based residuals for survival models. *Biometrika* **77:** 147-160.

Thibaudeau, Y. and Winkler, W. E. 2002. Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints. *Technical Report RRS2002/9.* US Bureau of the Census, Washington DC.

Titterington, D. M. and Sedransk, J. 1989. Imputation of missing values using density estimation. *Statistics & Probability Letters* **8:** 411-418.

Turnbull, B. W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38:** 290-295.

Vanichseni, S., Kitayaporn, D., Mastro, T. D., Mock, P., Raktham, S., Des Jarlai, D. C., Suharita, S., Srisuwanvilai, L., Young, N. L., Wasi, C., Subbarao, S., Heyward, W. L., Esparza, J. and Choopanya, K. 2001. Continued high HIV-1 incidence in a vaccine trial preparatory cohort of injecting drug users in Bangkok, Thailand. *AIDS* (London, England) **15(3):** 397-405.

Wang, C. Y., Wang, S., Gutierrez, R. G. and Carroll, R. J. 1998. Local linear regression for generalized linear models with missing data. *Annals of Statistics* **26:** 1028-1050.

Wei, G. C. G. and Tanner, M. A. 1991. Applications of Multiple Imputation to the Analysis of Censored Regression Data. *Biometrics* **47:** 1297-1309.

Whittemore, A. S. and Keller, **J. B**. 1986. Survival estimation using splines. *Biometrics* **42:** 495-506.

Zhang, P. 2003. Multiple Imputation: Theory and Method. *International Statistical Review* **71:** 581-592.

Zhang, W., Zhang, Y., Chaloner, K. and Stapleton, J. T. 2009. Imputation methods for doubly censored HIV data. *Journal of Statistical Computation and Simulation* **79:** 1245-1257.

Zio, M. D., Scanu, M., Coppola, L., Luzi, O. and Ponti, A. 2004. Bayesian Networks for Imputation. *Journal of the Royal Statistical Society, Series A* **167:** 309-322

Appendix A

Radiotherapy alone			Radio-and Chemotherapy		
(45,∞)	(25, 37]	(37,∞)	(8, 12]	(0, 5]	(30, 34]
(6, 10]	(46,∞)	(0, 5]	(0, 22]	(5, 8]	(13,∞)
(0, 7]	(26, 40]	(18,∞)	(24, 31]	(12, 20]	(10, 17]
(46,∞)	(46,∞)	(24,∞)	(17, 27]	(11,∞)	(8, 21]
(46,∞)	(27, 34]	(36,∞)	(17, 23]	(33, 40]	(4, 9]
(7, 16]	(36, 44]	(5, 11]	(24, 30]	(31,∞)	(11,∞)
(17,∞)	(46,∞)	(19, 35]	(16, 24]	(13, 39]	(14, 19]
(7, 14]	(36, 48]	(17, 25)	(13,∞)	(19, 32]	(4, 8]
(37, 44]	(37,∞)	(24,∞)	(11, 13]	(34,∞)	(34,∞)
(0, 8]	(40,∞)	(32,∞)	(16, 20]	(13,∞)	(30, 36]
(4, 11]	(17, 25]	(33,∞)	(18, 25]	(16, 24]	(18, 24]
(15,∞)	(46,∞)	(19, 26]	(17, 26]	(35,∞)	(16, 60]
(11, 15]	(11, 18]	(37,∞)	(32,∞)	(15, 22]	(35, 39]
(22,∞)	(38,∞)	(34,∞)	(23,∞)	(11, 17]	(21,∞)
(46,∞)	(5, 12]	(36,∞)	(44, 48]	(22, 32]	(11, 20]
(46,∞)			(14, 17]	(10, 35]	(48,∞)

Intervals (in months) of cosmetic deterioration (retraction) for early breast cancer

Appendix B

A program for analysis of failure time data with dependent interval censoring

Thu, 05/03/2007 - 16:08 — pukku

by Dianne Finkelstein and David Schoenfeld

depcen.exe is a program for estimating survival probabilities and probabilities of attending visits as described in the paper "Analysis of Failure Time Data with Dependent Interval Censoring" (Finkelstein D.M., Goggins W.B, and Schoenfeld D.A., Biometrics 2002 58:298-304). The program was implemented in Matlab and runs as a batch job from a DOS command prompt. The time to blood shedding data from the paper is also included. "interval_censr_data.zip" contains the data in .dat format and the .sas file required for setup. When using this data, please reference the article cited above.

- depcen.zip

interval_censr_data.zip

http://hedwig.mgh.harvard.edu/biostatistics/node/15

Appendix C

```
                       The SAS System    07:21 Wednesday, December 9, 2009  87

                     The FREQ Procedure

                                  Cumulative    Cumulative
  Obs    Frequency     Percent     Frequency      Percent
  ffffffffffffffffffffffffffffffffffffffffffffffffffffffff
    3           1        0.49            1         0.49
    4           1        0.49            2         0.98

  234           1        0.49          204       100.00




                       The SAS System    07:21 Wednesday, December 9, 2009  91

                     The FREQ Procedure

                                  Cumulative    Cumulative
  SEX    Frequency     Percent     Frequency      Percent
  ffffffffffffffffffffffffffffffffffffffffffffffffffffffff
  F            11        5.39           11         5.39
  M           193       94.61          204       100.00


                                   Cumulative    Cumulative
  RACE    Frequency      Percent     Frequency      Percent
  fffffffffffffffffffffffffffffffffffffffffffffffffffffffff
     1         178       87.25           178        87.25
     2          23       11.27           201        98.53
     3           3        1.47           204       100.00
                       The SAS System    07:21 Wednesday, December 9, 2009  97

                     The FREQ Procedure

                                     Cumulative    Cumulative
fcd4stat    Frequency      Percent     Frequency      Percent
fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
       0         111       54.41           111        54.41
       1          93       45.59           204       100.00


                                     Cumulative    Cumulative
lcd4stat    Frequency      Percent     Frequency      Percent
fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
       0         138       67.65           138        67.65
       1          66       32.35           204       100.00


                                     Cumulative    Cumulative
sheddind    Frequency      Percent     Frequency      Percent
fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
       0          88       43.14            88        43.14
       1         116       56.86           204       100.00


                                     Cumulative    Cumulative
blposind    Frequency      Percent     Frequency      Percent
fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
       0         174       85.29           174        85.29
       1          30       14.71           204       100.00
```

```
                                         Cumulative    Cumulative
urposind    Frequency     Percent     Frequency      Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
       0          88       43.14            88        43.14
       1         116       56.86           204       100.00


                                         Cumulative    Cumulative
deathcen    Frequency     Percent     Frequency      Percent
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
       0         154       75.49           154        75.49
       1          50       24.51           204       100.00
```

Appendix D

```
                                       The SAS System    07:21 Wednesday, December 9, 2009 120

                                     The CORR Procedure

            6  Variables:    FIRSTCD4 LASTCD4  BOFF     UOFF     BSURVTM  USURVTM


                                     Simple Statistics

 Variable           N          Mean       Std Dev           Sum       Minimum       Maximum

 FIRSTCD4         204     107.97549     115.65178         22027             0     568.00000
 LASTCD4          204      76.69118     111.95314         15645             0     652.00000
 BOFF             204      15.32353       5.23000          3126             0      28.00000
 UOFF             204      15.46078       5.34370          3154             0      31.00000
 BSURVTM          204      10.79412       6.17234          2202             0      25.00000
 USURVTM          204       6.70343       6.35362          1368             0      21.00000


                         Pearson Correlation Coefficients, N = 204
                                 Prob > |r| under H0: Rho=0

              FIRSTCD4       LASTCD4          BOFF          UOFF       BSURVTM       USURVTM

FIRSTCD4       1.00000       0.84649       0.28628       0.25951       0.36085       0.35217
                              <.0001        <.0001        0.0002        <.0001        <.0001

LASTCD4        0.84649       1.00000       0.15642       0.12979       0.25436       0.34932
                <.0001                      0.0255        0.0643        0.0002        <.0001

BOFF           0.28628       0.15642       1.00000       0.96073       0.73882       0.22772
                <.0001        0.0255                      <.0001        <.0001        0.0011

UOFF           0.25951       0.12979       0.96073       1.00000       0.72038       0.24744
                0.0002        0.0643        <.0001                      <.0001        0.0004

BSURVTM        0.36085       0.25436       0.73882       0.72038       1.00000       0.35181
                <.0001        0.0002        <.0001        <.0001                      <.0001

USURVTM        0.35217       0.34932       0.22772       0.24744       0.35181       1.00000
                <.0001        <.0001        0.0011        0.0004        <.0001
```

Appendix E

Analysis of subjects by CMV shedding and disease stage in both blood and urine.

Shedding in:		blood shedding	urine shedding
Early disease stage	Number of subjects	7	27
	Mean CD4+ count	143	196.7
Late disease stage	Number of subjects	15	41
	Mean CD4+ count	21.1	27.8
Total number of subjects		22	68
Mean CD4+ count		59.9	94.9

Appendix F

Analysis of subjects by CMV non-shedding and disease stage in both blood and urine.

Non-Shedding in:		blood	urine
Early disease stage	Number of subjects	86	66
	Mean CD4+ count	211.3	210
Late disease stage	Number of subjects	96	70
	Mean CD4+ count	26.4	24.4
Total number of subjects		182	136
Mean CD4+ count		113.8	114.5

OFFICE OF THE DEPUTY VICE-CHANCELLOR:
ACADEMIC AFFAIRS AND RESEARCH
Private Bag X1314, Alice 5700
Tel: 04060 22403
Fax: 0866282944
tsnyders@ufh.ac.za

University of Fort Hare
Together in Excellence

REC-270710-028

Application for clearance from the University of Fort Hare's Ethics Committee

Project title: **A COX PROPORTIONAL HAZARDS MODEL FOR MID-POINT IMPUTED INTERVAL-CENSORED DATA**

Chief Researcher: Arnold Rumosa Gwaze

Supervisor: Professor J Tyler

Date of application: 6 December 2010

Having consulted the Dean of Research, I hereby grant permission to conduct the research.

Professor J R Midgley
Deputy Vice-Chancellor
Chairperson of the interim Ethics Committee

7 December 2010