MODELLING RECURRENT EPISODES OF PERITONITIS AMONG PATIENTS WHO ARE IN PERITONEAL DIALYSIS AT PIETERSBURG PROVINCIAL HOSPITAL, LIMPOPO PROVINCE, SOUTH AFRICA

by

THEMBHANI HLAYISANI CHAVALALA

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER of SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Prof. Y.G KIFLE
CO-SUPERVISOR: Dr. R.A TAMAYO
(Pietersburg Provincial Hospital)

2019

Declaration

I, **Thembhani Hlayisani Chavalala**, declare that the dissertation which is hereby submitted for the qualification of Master of Science in Statistics at the University of Limpopo, titled: Modelling Recurrent Episodes of Peritonitis among Patients who are in Peritoneal Dialysis at Pietersburg Provincial Hospital, Limpopo Province, South Africa, is my own work and all cited and quoted work has been fully referenced. I, further declare that this work has not been submitted before, either by me or anyone at any other institution in South Africa, or outside of South Africa.

Signature: .....

Date: 03 July 2019

Dedication

I thank God for the success of this project. I dedicate this work to my son, Andziso Kgothalo Chavalala and my late mother, Julia Nurse Lephaka. You will always have my heart.

Acknowledgements

I would like to offer my sincere gratitude to my supervisor, Prof. Y.G. Kifle and co-supervisor, Dr. RA Tamayo Isla who have offered full support, valuable advice and guidance throughout the project. Messrs Maluleke H and Dikgale RP, you are acknowledged for the vital role you have played. I would also like express my gratitude to my family and friends for believing in me and encouraging me throughout this research. It is because of your day and night prayers that I didn't give up during the bad seasons of this project. Moreover, this project would not have been became a success without the financial support provided by the VLIR-OUC Programme. Most importantly, I thank my Creator for blessing me with these people and opportunity.

Abstract

Recurrent peritonitis is a major problem of peritoneal dialysis (PD) due to its association with technique failure in the dialysis process. The literature on peritonitis focused only on investigating major risk factors associated with the first episode of peritonitis. However, this dissertation investigates factors associated to multiple episodes of peritonitis, to a maximum of 6 episodes. The correlation of recurrent episodes of a patient is considered.

The univariate counting process, stratified, gap-time and marginal hazard regression models are applied to select the significant covariates to the multivariate regression hazard models. Regression coefficient for covariates are found to be statistically significant at 5% level. The application of Akaike information criterion (AIC) and Schwarz bayesian criterion (SBC) assisted to filter out the best method which is the stratified regression hazard model. The major risk factors associated with recurrent episodes of peritonitis are examined from the selected good fitting model.

In conclusion, the selected model identified two independent risk factors to be significantly associated with recurrent episodes of peritonitis: marital status and glomerular filtration rate. Two categories of marital status, divorce and widower are the significant factors compared to married patients (when taking married patients as the reference category).

Keywords: Peritoneal dialysis, peritonitis, recurrent episodes, survival analysis.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criterion
AFT	Accelerated Failure Time
BMI	Body Mass index
APD	Automated Peritoneal Dialysis
CAPD	Continuous Ambulatory Peritoneal Dialysis
CCPD	Continuous Cycling Peritoneal Dialysis
PKDC	Polokwane Kidney Dialysis center
ESRD	End-Stage Renal Disease
HD	Haemodialysis
HR	Hazard Ratio
ID	Identity
KM	Kaplan-Meier
NA	Nelson-Aalan
PD	Peritoneal Dialysis
RRF	Residual Renal Function
SA	South Africa
SAS	Statistical Analysis System
SBC	Schwarz Bayesian Criterion
STD	Standard Deviation
SSE	Sum Square Error
US	United State

Chapter 1

INTRODUCTION

1.1 Introduction

Section 1.1 of this chapter gives a brief introduction about peritonitis, survival analysis and the study background. Section 1.2 focuses on the research problem. Section 1.3 discusses the purpose of the study through the study motivation, aim and objectives. Whereas the last three Sections, 1.4, 1.5 and 1.6, outline the scientific contributions, data structure and structure of the study, respectively.

1.1.1 Peritonitis

Kidneys are body organs which filter unneeded water, waste products and other body chemicals from the person's blood. When the kidneys fails to perform its duties, a person develops a condition called kidney failure. Kidney failure also known as renal failure is defined as a situation where the body organs fails to remove waste product and extra fluids from the blood. This condition can cause sickness

since the waste product, chemicals and extra body fluids are going to build up in the blood. However, there are two ways of treating kidney failure, which are dialysis and kidney transplant. Kidney transplant is a treatment where by the damaged kidneys is removed and replaced by healthy kidneys. Dialysis is a machine treatment which removes the unneeded fluids in the blood. Moreover, dialysis is divided into hemodialysis and peritoneal dialysis.

Peritoneal dialysis (PD) is a kidney failure treatment which uses the lining of the abdomen called peritoneum, to filter and remove waste products in the blood. However, before the patient can start with this treatment, a thin tube called catheter must be inserted inside the patient's belly as shown in Figure 1.1. Catheter is a device which works as a transporter. It transport the dialysis solution in and out side the belly. When it is done performing its duties, that is, when the dialysis solution bag is empty, the patient can remove it and continue doing his or her daily activities. Dialysis solution also known as the dialysate soaks up unneeded wa-

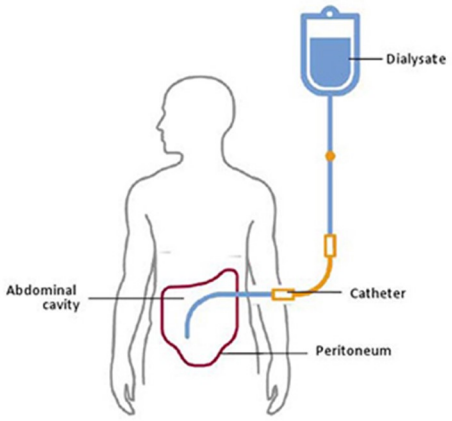

Figure 1.1: Example of peritoneal dialysis treatment

ter, waste products and other body chemicals from the patient's body into the PD bags. The used dialysis solution can be drained into the PD bag and replaced with a fresh bag of dialysate. The process of draining the used dialysis solution and replace it with a fresh dialysate is called exchange. PD patients are usually trained to perform the exchange process on their own. Most of the patients get discharged

after acquiring the skills of performing the exchange process on their own. Due to this reason, many PD patients perform the exchange process at home without the doctor's supervision.

PD is divided into two parts, namely: continuous ambulatory peritoneal dialysis (CAPD) and automated peritoneal dialysis (APD). The CAPD is well-known treatment modality in end-stage renal disease (ESRD) patients. It does not require the machine to filter the waste products and extra body fluids from the patient's blood. Whereas, APD is kidney failure treatment which uses the machine called cycler to drain and fill the patient's belly. Due to the need for improving the PD patient's quality of life, APD has therefore became a daily home treatment with automated nightly exchanges (Roberto et al., 2007). Nonetheless, this treatment has its own shortcomings.

The most popular shortcoming of PD is that it can lead to an infection called peritonitis. Peritonitis is a serious abdominal infection which occurs when the exit-site of the catheter becomes infected. Moreover, peritonitis can also occurs when the catheter becomes contaminated as the patient connect or disconnect it from the dialysis solution bag. Peritonitis is one of the infections which can occur several times on the same PD patient. That is, peritonitis can reoccur after being treated. When this happens, that particular patient is said to have experienced recurrent (multiple) episodes of peritonitis. A person with peritonitis mostly experience one of the following symptoms:

- Abdominal pains,
- fever,
- vomiting,
- pains around the catheter, and
- cloudiness in the used dialysis solution and a catheter cuff that pushes out of the body.

Peritonitis is a life threatening infection and therefore it is essential to report these symptoms to the doctor so that they can start treating it as early as possible. The instructions given to the PD patients in order to avoid peritonitis includes:

- Cleaning the exit-site of the catheter,
- washing the hands every time you need to perform exchange,
- wearing surgical mask when performing exchange, and
- inspecting each dialysis solution bag for the signs of contamination.

1.1.2 Survival analysis

Survival analysis is a statistical technique appropriate for analysing data in which the primary interest is time that takes an object to experience the designated event of interest (Kleinbaum and Klein, 2010). This is unlike in logistic regression where the primary interest is how the risk factors were associated with whether the object has experienced the designated event or not (absence or presence of a particular event). In survival analysis the designated event of interest is called failure (Stevenson and EpiCentre, 2009). That is, the occurrence of the designated event is called failure even though the event itself sometimes is a success. For example, when the patient recovers from heart transplant, it is called failure even though the patient has succeeded. Furthermore, the time it takes an object to fail is called failure time, event time or survival time.

The exact failure time is only known for the object that have experienced the event of interest before the end of the follow-up period (Stevenson and EpiCentre, 2009). Objects that did not experience the designated event during the follow-up period are said to be censored. Censoring is the key analytical problem which distinguishes survival analysis and the other ordinary statistical techniques. However, there are different types of censoring, namely, right, left and interval censoring.

Right censoring is the most commonly encountered form of censoring (Liu, 2012). It occurs when only the lower bound of the failure time of the object is known.

Survival analysis of recurrent episodes considers a situation where an individual may experience more than one designated event of interest over the follow-up period. There are many statistical techniques that can be utilised to handle analysis of this nature. However, this study utilises the stratified, gap-time, marginal and the counting process techniques. Irrespective of which technique is being utilised, the variance of the estimated partial likelihood regression coefficients should be adjusted for the possible correlation among recurrent episodes within the same patient. The sandwich robust variance estimator is the most popular and widely used estimator for adjusting the variances of the partial likelihood estimated regression coefficients (Childers, 2015). Therefore, this study also utilises it to avoid possible correction among the multiple episodes within the same patient.

Gap-time scale model is the most often used in studying a multiple events rate as a function of time since the last event (Duchateau et al., 2003). In gap-time model, a patient moves to the K^{th} stratum immediately after the $(K-1)$ recurrence time and remains there until the K^{th} episode occurs or until the patient is censored. In general, a patient with K episodes contributes $K+1$ observations. Counting process approach uses the standard Cox proportional hazard model. However, the various time interval on the same patient are used in the formation of the likelihood function. These time interval are considered as independent intervals from different patients, even though they are coming from the same patients. Patients remains at risk set of episodes until the last follow-up visit.

Stratified approach uses the same approach used for the counting process, except that it uses the stratified Cox model rather than the standard Cox model. In stratified model, the time until the occurrence of the first episode affects the composition of the risk set for later episodes. The marginal approach uses the Cox proportional hazard model while leaving the nature of dependence among correlated failure

times completely unidentified. Patients are considered to be at risk of all episodes, regardless of how many episodes each patient have actually experienced.

1.1.3 Research background

Despite the fact that PD is a well-established treatment for kidney failure or ESRD, peritonitis remains the major problem for the wide utilisation of PD. This is due to its association with high morbidity in patients on long-term CAPD, technique failure and the need to transfer PD patients to haemodialysis (HD), which is an expensive treatment of kidney failure (Martin et al. (2011) and Mashiloane et al. (2008)). Peritonitis makes continuing of PD impossible even when the catheter is being used to protect against the spread of bacteria. In South Africa (SA), Mashiloane et al. (2008) indicated that peritonitis is the major limiting factor for CAPD and that the increased frequency of peritonitis was influenced by the poor socio-economic conditions, age and diabetes. Ikabu et al. (2007) stated that ESRD is a serious burden for both patients and health care professionals of SA, whereas Isla et al. (2014) indicated that the prevalence of ESRD continues to increase world-wide including in many developing countries. Most of the renal failure patients treated with PD have shown to have a lower risk of death early during the course of ESRD and that there is no significant difference in the risk of death among patients with ESRD incidents treated with HD or PD (Mehrotra et al., 2011).

The most common problem in PD patients who had peritonitis is that peritonitis can reoccur for the second or more times after being treated. The condition known as a recurrent episodes of peritonitis is a major complication of CAPD and it remains the leading cause of treatment technique failure and catheter loss in the PD patients (Vonesh (1985) and Nieto-Ríos et al. (2014)). Nieto-Ríos et al. (2014) reported that independently of other factors patients with recurrent episodes of peritonitis have higher chances of death. Mashiloane et al. (2008) argued that black patients tends to be the ones experiencing recurrent episodes of peritonitis as compared to white

patients.

1.2 Research problem

Peritonitis has become a life threat to PD patients and a major complication of treatment failure. When the ongoing of PD becomes impossible due to peritonitis, patients need to be transferred to hemodialysis (HM). HM is an expensive kidney failure treatment as compared to PD. Patients who cannot afford HM end-up losing their lives. When people die, the country face a serious problems such as losing qualified and economically active people who are the breadwinners, leaders, and helpers of the country's next generation. The number of people dying each and every year can increase the risk of inflation in the country as it reduces the number of workers. Peritonitis is reported to be influenced by poor socio-economic conditions, age, diabetes and the poorer population which the majority of it is the black race and have historically suffered a low standard of health care. In order to address the above problem, four types of recurrent regression hazard model are employed to investigate the risk factors influencing peritonitis. However, factors from the better fitting model are considered as the major risk factors associated with peritonitis.

There are studies conducted on peritonitis, however, most of them focused on investigating factors associated with the first episode of peritonitis, (Martin et al. (2011) and Fan et al. (2014)). Since peritonitis is one of the infections which can occur several times in the same patient, conducting a study looking at the first episode could lead to misleading conclusions. This is due to the reason that some useful information such as whether the occurrence of the first episode influences the occurrence of the other episodes would be ignored. It is essential for researchers to avoid loosing information during the study process and start addressing this infection as a recurrent events problem. Using the complete information will improve

the quality of the results and correct the mistakes committed. This study focuses on the recurrent episodes of peritonitis rather than looking at the first episodes of peritonitis as have been done by previous researchers.

Dataset for survival analysis includes the censoring variable. The censoring variable is a variable which records the occurrence of the designated experience of interest (event) and the termination of the survival process. The variable is recorded as binary, recording those who got the event of interest (peritonitis) and those who are lost during the follow-up period, withdraws during the study or study end before getting peritonitis (censored). Due to this censoring variable, survival analysis tends to be more similar to some conventional statistical perspectives on qualitative outcome data, such as the logistic regression and the probit model. These statistical techniques can be applied to examine the occurrence of a particular event of interest by comparing the status of the individual at the beginning and at the end of the follow-up period (Liu, 2012). However, these techniques does not consider the time at which the patient experienced peritonitis or the time at which the patients got censored, and therefore they do not have the characteristics of describing the time to event process (Liu, 2012). Failing to posses these characteristics can cause damage to the quality of the analytic results, thereby generating misleading conclusions.

In many of the previous studies, the logistic regression model has been utilised to analyse the time-to-event data (Gray et al., 2013). Logistic regression model does not consider the time at which the event occurs and therefore, disregards the length of the survival process. Therefore, this study focuses on utilising the appropriate recurrent survival modelling techniques in order to address peritonitis as a recurrent infection.

1.3 Purpose of the study

1.3.1 Motivation

There is insufficient information available on the factors associated with the recurrent episodes of peritonitis, since most of the previous studies on peritonitis focused much on the first episode of peritonitis. Therefore, this study might add to the little information available by assessing the factors associated with the recurrent episodes of peritonitis. As per the results obtained by Martin et al. (2011); age, black race, diabetes and congestive heart failure seems to be the significant risk factors. However, they suggested that more studies looking at factors influencing the peritonitis incidence should be conducted in order to improve the PD outcome. Hence, the current study aims at revealing the major risk factors associated with the recurrent episodes of peritonitis in order to improve the PD outcome. The study conducted by Fan et al. (2014) revealed that older age, male, lower educational level and hypoalbuminemia at initiation of PD are associated with the first episode of peritonitis. They went on to indicate that these results might be useful in identifying patients starting PD treatment who are at high risk for the first episode and on how to improve CAPD outcome. Thus, this study seek to disclose more on peritonitis but focusing on recurrent episode rather than the first episode.

1.3.2 Aim

The aim of the study is to investigate risk factors associated with the recurrent episodes of peritonitis.

1.3.3 Objectives

The objectives of this study are to:

1. Evaluate four models: counting process, gap-time, marginal and stratified hazard model.
2. Apply peritonitis dataset to fit counting process, gap-time, marginal and stratified models.
3. Compare the four models mentioned in point 2 above, to select the best model to conduct the analysis.
4. Identify possible social and biological risk factors associated with recurrent episodes of peritonitis.
5. Assess the rate at which kidney patients who are on PD get exposed to recurrent episodes of peritonitis.

1.4 Scientific contribution

The findings of this study could assist the Department of Health in Limpopo province mainly by providing useful information about various potential risk factors associated to the recurrent episode of peritonitis in kidney patients who are on PD. Moreover, the findings of this study could also be used in educating PD patients about controllable social and biological risk factors of peritonitis. Furthermore, as this study apply and compare various recurrent survival analysis techniques, the outcome will give a direction to statisticians as which appropriate technique(s) to apply in a situation where there are recurrent episodes.

1.5 Data structure

A sub-sample data of five PD patients in Table 1.1 demonstrate the general data layout required for modelling the time to recurrent episodes of peritonitis. The first column contains the patient's identification (ID) number, the second column

indicate the order of visits/stratum, the third and forth columns give the time start and time stop (in days) for each patient per visit/stratum. The time start is the day the patient entered the study while the time stop is the day the patient experienced peritonitis or censorship. The fifth column gives the gap time between the time start and time stop for each patient per visit. The Sixth column contains the censoring status taking the value one (status=1) if peritonitis and value zero (status=0) if censored. The seventh, eighth and ninth columns give the gender, age at baseline and race of the patients, respectively.

The data set consist of the first six visits for each PD patient, however only few patients experienced the targeted six episodes of peritonitis during these visits. A patient who did not experience all six episodes will have some missing values. For instances, it can be seen from Table 1.1 that a patient with ID number 1 began the follow-up at time 0 and remained at risk until day 1008 yet did not experience even a single episode. This imply that there are some patients who were free to peritonitis during all the six visits. Patient with ID number 18 experienced four episodes from the first four visits and did not get the event at the last two visits. The patient (ID number 143) was followed for 1190 days yet experienced the six targeted recurrent episodes of peritonitis. The last patient (ID number 152) was in the follow-up period for 84 days yet had one episode during the first visit.

Despite the number of visits recorded per patient, some patients got censored before under going the six targeted visits. This can be seen by checking the gap time variable which records the gap between the day the patient entered the study and the day the patient experienced peritonitis or censored. The value of zero in the gap time variable indicates that the time start and time stop are the same and therefore the patient was out of the study during the corresponding follow-up visit. The time to peritonitis is measured in days from the beginning of follow-up until the occurrence of peritonitis or until the patient is censored.

Table 1.1: Time to recurrent episodes of peritonitis data layout

ID	Visit	Tstart	Tstop	Gap	Status	Sex	Age	Race	...
1	1	0	1008	1008	0	Male	39	Black	...
1	2	1008	1008	0	0	Male	39	Black	...
1	3	1008	1008	0	0	Male	39	Black	...
1	4	1008	1008	0	0	Male	39	Black	...
1	5	1008	1008	0	0	Male	39	Black	...
1	6	1008	1008	0	0	Male	39	Black	...
..	..	..	..	..	..	..	..	..	...
..	..	..	..	..	..	..	..	..	...
..	..	..	..	..	..	..	..	..	...
18	1	0	464	464	1	Male	45	Indian	...
18	2	464	956	492	1	Male	45	Indian	...
18	3	956	1130	174	1	Male	45	Indian	...
18	4	1130	1401	271	1	Male	45	Indian	...
18	5	1401	1727	326	0	Male	45	Indian	...
18	6	1727	1727	0	0	Male	45	Indian	...
..	..	..	..	..	..	..	..	..	...
..	..	..	..	..	..	..	..	..	...
..	..	..	..	..	..	..	..	..	...
143	1	0	21	21	1	Female	44	Black	...
143	2	21	61	40	1	Female	44	Black	...
143	3	61	607	546	1	Female	44	Black	...
143	4	607	688	81	1	Female	44	Black	...
143	5	688	965	277	1	Female	44	Black	...
143	6	965	1190	255	1	Female	44	Black	...
..	..	..	..	..	..	..	..	..	...
..	..	..	..	..	..	..	..	..	...
..	..	..	..	..	..	..	..	..	...
151	1	0	10	10	1	Male	51	White	...
151	2	10	64	50	1	Male	51	White	...
151	3	64	64	0	0	Male	51	White	...
151	4	64	64	0	0	Male	51	White	...
151	5	64	64	0	0	Male	51	White	...
151	6	64	64	0	0	Male	51	White	...
152	1	0	23	23	1	Male	17	Black	...
152	2	23	84	61	0	Male	17	Black	...
152	3	84	84	0	0	Male	17	Black	...
152	4	84	84	0	0	Male	17	Black	...
152	5	84	84	0	0	Male	17	Black	...
152	6	84	84	0	0	Male	17	Black	...

1.6 Structure of the study

The study is divided into five chapters. Following this introductory chapter, chapter two present literature review on modelling techniques necessary to gain insight of the data analysis. Chapter three provide the theoretical aspects and application of the survival analysis methods on time to recurrent episodes of peritonitis. The fourth chapter present the results and discussion of the time to recurrent episodes of peritonitis. Finally, the last chapter concludes the study by providing key findings, recommendations and suggestions for further studies.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the work done on peritonitis by other researchers. It looks at the modelling techniques and strategies the researchers used to get the results of their studies. Following this introductory Section 2.1, Section 2.2 looks at the factors associated with peritonitis. Section 2.3 reviews the modelling techniques used to produce the analytic results. Finally, Section 2.4 will summarise the chapter.

2.2 Factors associated with peritonitis

Fan et al. (2014) investigated the risk factors associated with the first episode of peritonitis in the Southern Chinese CAPD patients. Their study revealed that when looking at gender, men are associated with the high risk of the first episode of peritonitis when compared to women. These results does not correspond with the

one found by Fried et al. (1996) in the study of determining whether peritonitis influences mortality. Fried et al. (1996) indicated that when checking the survival of gender, men were significantly lower as compared to women. The significance of the female gender as a risk factor influencing peritonitis was confirmed again by Kotsanas et al. (2007).

In Brazil, Martin et al. (2011) conducted a study to identify the risk factors that influence the first episode of peritonitis. Educational level was found to be the strong risk factor associated with the first peritonitis episode. The association was independent of socio-economic conditions, PD mortality and commodities. Their findings are equivalent to the findings established by Fan et al. (2014), even though they indicated that the significance of educational level in their study was independent of hemoglobin and potassium.

The study by Nessim et al. (2009) on the impact of age on peritonitis risk of PD patients was motivated by the increasing number of elderly patients reaching ESRD. The variable age was categorised as older if the patient's age is at leat 70 years. Their PD initiation was divided into two eras, 1996 to 2000 and 2001 to 2005. The study revealed that older age patients were independently associated with peritonitis among patients initiated PD between 1996 and 2000. Moreover, it was found to be not associated with the ones initiated between 2001 and 2005. This association was confirmed again after 5 years by Fan et al. (2014). In the study of analysing the clinical and bacteriological factors associated with the shock and mortality in patients with secondary generalised peritonitis, Riché et al. (2009) indicated that the age over 65 was found to be the independent risk factor. However, the results of these studies contradicts the finding of Port et al. (1992) who indicated that significantly higher risk of peritonitis and technique failure was observed for younger patients.

Keleş et al. (2010) addressed the issue of PD through analysing the risk factors associated with peritonitis in PD patients at Northeast Anatolia. In their study they

tried to compare PD patients who never experienced peritonitis with PD patients who experienced at least one episode of peritonitis. The outcomes of their study reflected that hypoalbuminemia, constipation, placement of catheter through surgery and amyloidosis are the factors increasing the risk of peritonitis in PD patients. Hypoalbuminemia was confirmed again to be a significant risk factor of the first episode of peritonitis in the study conducted by Fan et al. (2014).

Isla et al. (2016) examined the prevailing causes and predictors of mortality among predominantly rural dwelling ESRD patients in SA. Their study indicated that there was no difference in age, gender, race and predominant areas of dwelling. However, they found statistical significant difference in the types of housing, with more CAPD patients dwelling in rural formal houses and in the survival times between HD and CAPD patients. The conclusion drawn in the study is that poor access to health care facilities plays a vital role in infection-related mortality. Gray et al. (2013) compared PD patients characteristics and outcomes in the rural and urban areas of Australia. Their study revealed that PD technique failure rates are low in rural areas than in the urban areas.

Okayama et al. (2012) conducted a study titled ”aging is an important risk factor for peritoneal dialysis-associated peritonitis” in order to determine the risk factors associated with peritonitis. The event of interest was peritonitis, while sex, age, diabetes mellitus and several laboratory values were some of the studied variables. The study revealed that rather than diabetes mellitus, aging was the important risk factor of PD associated with peritonitis.These results contradict the finding of (Han et al. (2007), Chow et al. (2005), Oo et al. (2005) and Golper et al. (1996)) whom indicated that diabetic status was the independent risk factor for peritonitis in CAPD patients.

Nieto-Ríos et al. (2014) studied the rate of CAPD-related peritonitis in a cohort study of patients followed for 27 years at a single PD center. The study revealed that there was no significant changes in the 27 years of follow-up, i.e., the study

showed that the rate of peritonitis was stable for the whole 27 years follow-up period. However they also reported that access to health care services and the distance travelled to the PD center turns to be some of the socio-economic factors that could influence the rate of peritonitis for PD initiated patients. The issue of access to healthcare facilities is in agreement with the argument highlighted by Isla et al. (2016).

The study of comparing peritonitis rate between CAPD and continuous cycling peritoneal dialysis (CCPD) was conducted in United States (US) by Oo et al. (2005). The study revealed that the risk of peritonitis for CAPD patients was lower as compared to the risk of CCPD patients. Black race was found to be the significant risk factor associated with peritonitis. The significance of black race in their study was confirming the results found by Port et al. (1992). Even though these findings were validated by the study conducted in Brazil by Martin et al. (2011), they contradict the findings from the study of Fried et al. (1996) who indicated that peritonitis was a risk factor only in white PD patients.

Isla et al. (2014) assessed the outcome of patients treated with CAPD in Limpopo, South Africa. The rate of peritonitis and the factors influencing peritonitis were also investigated. Out of 152 patients who entered the study, 71 (46.7 percent) of them reached the composite outcome of death or technique failure. The overall number of infections reported during the study period were 210 with the peritonitis rate of 0.82 per year. At the end of the study period only 66 (43.4 percent) of patients were still active on CAPD. It is therefore correct to conclude that more than half of the studied patients were censored. This censoring was due to reasons such as death, technique failure and transferred to other CAPD centers. Hemoglobin, serum albumin, body mass index (BMI) and experiencing more than one episode of peritonitis were reported to be the factors identified to predict the composite outcome.

Zent et al. (1994) evaluated the specified biomedical, socio-economic, and psy-

chosocial criteria as predictors of therapeutic success to improve patients selection process for CAPD in developing countries. In their study they also investigated the presence of the relationship between the episodes of peritonitis and other exit-site infections. This study revealed that the rates of peritonitis were high, especially in black race. They went on and point out that age, black race and diabetes were the factors connected with the increased peritonitis rates. The connections of these factors are in agreement with the results found in Brazil by Martin et al. (2011). They concluded that the connections of these factors with the high rates of peritonitis could have serious implications on how to select patients for CAPD.

2.3 Techniques used to model time to peritonitis

Schneider et al. (2009) examined the prognostic factors in the critically ill patients suffering from secondary peritonitis. The changes of survival time during the window period, septic patients with or without peritonitis were evaluated using the kaplan-Meier estimator, log-rank test and the generalised Wilcoxon test. The Cox proportional regression model was employed to assess the association of the variables and the survival time. The Schoenfeld residuals method was used to plot the residuals for assessing the form of the relationship between survival time and patient variables and also used to check the Cox proportional hazard model assumptions.

Feng et al. (2016) conducted a study to compare the prognosis of early onset peritonitis and non-early onset peritonitis in PD patients. The continuous variables between groups were compared using student's t-test while the categorical variables were compared using the Pearson chi-square test. The presence of normal distribution in continuous variables were tested using the Kolmogorov-Smirnov test. Survival curves were plotted using the Kaplan-Meier estimator and tested for significance difference using the log-rank test. Variables with p-value less than 10%

in the univariate Cox regression model were subjected to the multivariate Cox regression model. In the multivariate model variables were considered statistically significant if p-value is less than 5%.

Isla et al. (2014) assessed the peritonitis rate and the causes of peritonitis for patients treated with CAPD in Limpopo, SA. The analytic results were generated from survival analysis techniques. Survival curves of patients were obtained and tested for statistical significant difference using the Kaplan-Meier estimator and the log-rank test. The univariate Cox regression model was employed to select the significant variables that can be included in the multivariate regression. The adequacy of the multivariate Cox proportional hazard model was assessed using the Hosmer-Lemeshow test and variables were considered to be significant at the multivariate method if p-value is less than 5%.

Han et al. (2007) evaluated the effect of residual renal function (RRF) on the development of peritonitis in patients treated with CAPD. The patients were grouped as peritonitis and peritonitis free. These groups were compared if they are the same or not using the student's t-test if variable is continuous and chi-square test for categorical variables. The survival of these patients was examined using the Kaplan-Meier estimator and the log-rank test. The risk factors associated with peritonitis were also investigated and this was done using the multivariate Cox proportional regression hazard model.

Isla et al. (2016) conducted a study to identify the existing causes and predictors of mortality among ESRD patients who are mainly dwelling in rural areas of Limpopo, SA. The median time to survival of patients was determined in the study through the kaplan-Meier (product-limit) method. The survival curves generated from the product-limit method were assessed for the existing significant difference between using the log-rank test. Variables were selected to the multivariate analysis if there were statistical significant at 25% level of significance in the univariate Cox regression Model. Variables entered to the multivariate Cox proportional hazard model

were considered to be statistical significant at 5% level of significance.

Okayama et al. (2012) investigated the risk factors influencing peritonitis. The significance of the risk factors was evaluated using the Multivariate Cox proportional hazard model. The groups of patients with peritonitis and those without peritonitis were compared using the non-paired t-test and the chi-square test depending on whether the variables are continuous or categorical. The Peto log-rank test was employed to compare the difference in the occurrence rate of PD related peritonitis between patients with 65 years of age or more and with less than 65 years of age.

Keleş et al. (2010) conducted a study to identify the risk factors associated with peritonitis in PD patients. The continuous variables were compared using the student's t-test, while the categorical variables were compared using the chi-square test. Variables with p-value less than 20% in the univariate Cox regression model were considered to be statistical significant and were subjected to the multivariate Cox proportional hazard model. The backward elimination method was utilised in the Cox proportional model and the level of significance was considered to be 5%.

Barone et al. (2012) examined indices associated with peritonitis between CAPD and APD patients. The weighted t-test was utilised to compare the cumulative peritonitis rate among patients treated with CAPD and APD. The Kaplan-Meier analysis was employed to determine the probability of remaining free peritonitis from all peritonitis episodes and also used to calculate time to first peritonitis among the groups. The significance difference of the survival curves was compared through the log-rank test. The proportion of patients with peritonitis among these groups was assessed using the chi-square test and was considered statistical significant if p-value is less than 5%.

Rudnicki et al. (2010) investigated the risk factors for PD associated with the infection peritonitis. The significance of the continuous variables was tested using the unpaired two tailed t-test or the two tailed Mann-Whitney U-test, while the categorical variables were examined through the chi-square test. The survival curves

of patients using and not using oral vitamin D were generated using the Kaplan-Meier estimator and tested for significant difference using the log-rank test. The risk factors associated with peritonitis were investigated using the univariate Cox proportional hazard regression model.

2.4 Summary

This chapter has evaluated the literature on peritonitis. The evaluation of the models show that survival analysis techniques, such as, Kaplan-Meier estimator, log-rank test, and Cox proportional hazard regression model can be utilised to model the risk factors associated with peritonitis. It is also presented that peritonitis is associated with social and biological factors. These facts has assisted in selection of the appropriate methods and the variables to consider in the analysis.

The literature presented on this chapter depicts that most of the conducted studies about peritonitis utilised most of the techniques of ordinary survival analysis. To be more specifically, majority of the studies assessed the risk factors associated with the first episode of peritonitis. Peritonitis is one of the infections which can reoccur after being treated. Thus, it is very vital to start addressing peritonitis as a recurrent infection. Techniques of recurrent survival analysis are to be employed in this study to investigate the major risk factors associated with recurrent episodes of peritonitis.

Chapter 3

METHODOLOGY

3.1 Introduction

This chapter discusses the statistical data analysis techniques which was used to carry out the results of time to recurrent episodes of peritonitis in the forthcoming chapter. Survival analysis of recurrent episodes need the background of the ordinary survival analysis techniques. Thus, the usual survival analysis techniques are presented first as the foundation of building to recurrent survival analysis techniques.

3.2 Research design and data collection

The study presents analysis which is carried out using the prospective dataset collected on PD patients at Polokwane Kidney and Dialysis Center (PKDC) of the Pietersburg Provincial Hospital in Limpopo Province, South Africa. The data was

gathered from January 2008 to December 2012 and all patients were evaluated on a monthly basis at the PKDC for peritonitis. The interest of the study is on patients with the recurrent episodes of peritonitis (event of interest) whom were followed from 2008 to 2012 (follow-up period). For data management and analysis purposes, Statistical Analysis System (SAS) softwares was employed.

3.3 Survival data

Survival process is described the length of time for which a subject persists before the occurrence of a particular event. For example, in this study PD patients were followed for five years on monthly basis until the occurrence of peritonitis. Therefore it is important to describe the follow-up period and the event of interest in survival data. The most important feature of survival data is censoring, which is defined as the incomplete survival time status for some of the followed patients. For example, not all PD patients experienced recurrent episodes of peritonitis during the follow-up period and therefore, their true survival time status would not be known (the patients would be censored).

3.4 Survival analysis

Survival analysis is a collection of statistical methods for analysing data where the outcome variable of interest is time until the occurrence of an event of interest. The event of interest can be any designated experience of interest that may happen to an individual. For example, the occurrence of more than one episode of the infection called peritonitis is considered as an event of interest in this study. The time to an event can be measured in days, weeks, months, or years from the beginning of follow-up until the occurrence of an event. However, PD patients are followed in a monthly basis in this study.

In survival analysis the term failure, usually refers to the occurrence of the event even though sometimes the event of interest is a success. Example, recovery from heart transplant can be considered as the event of interest in survival data. The time variable is defined as the survival time because it gives the length of time taken for a failure to occur (Stevenson and EpiCentre, 2009). When each patient can experience more than one episode, the episode of interest occurs repeatedly in the same subject, the analysis is known as recurrent survival analysis or survival analysis of multiple events.

3.5 Censoring

In survival analysis there exist a key analytic problem known as censoring. This happens due to the reason that not all the followed patients will experience the designated event of interest. Observations are said to be censored when their true survival time is incomplete. That is, some information about the person's true survival time is known, however, the exact survival time is not known. In survival data, censoring frequently happens for many reasons such as:

- Patient got peritonitis before the study begins
- Patient do not experience peritonitis before the study end
- Patient get lost to follow-up during the study period due to death
- Patient withdraws from the study because of migration (moving from one hospital to the other)

Censoring is an important type of missing data in survival analysis, and it is usually required in order to avoid bias when it happens randomly and be noninformative (Liu, 2012). Censoring is divided into several specific types, namely: right, left, and interval censoring.

Patients are said to be right censored if it is known that the event of interest oc-

curred some time after the recorded follow-up period. Right censoring is the most commonly encountered form of censoring and this study will use the data set where subjects are right censored. For analytic convenience, descriptions of right censoring are often based on the assumption that an individual's censored time is independent of the actual survival time, thereby making the right censoring noninformative (Liu, 2012).

3.5.1 Mechanisms of right censoring

There are three types of right censoring, namely: Type I or fixed right censoring, Type II right censoring and Type III or random right censoring.

In Type I right censoring, each observation has a fixed censoring time, in such a way that a specific follow-up period is designed with a starting date and an ending date. In most cases, not the whole population would experience the event of interest during the specified follow-up interval. For those who will survive to the endpoint, the only available information will be that their exact survival time is located to the right of the endpoint of the follow-up period, denoted by $T > C$, where T is the failure time and C is the fixed censored time.

Type II right censoring is defined as a situation in which a fixed number of failures is targeted for a certain study. In Type II right censoring, the study terminate automatically when the targeted number of failures is observed and all the individuals whose survival time are greater than the time of termination are considered to be right censored, denoted by $T_i > T_r$, where T_i is the ordered lifetime for $i = 1, 2, 3, , ..., n$ and T_r is the lifetime of the targeted r-th failure.

Right censoring that happens randomly at any time during the follow-up period is referred to as random right censoring. Statistically, time for random censoring can be described by the random variable C_i (where i indicates variation in C among randomly censored observations), which is generally assumed to be independent

of lifetime T_i. For a sample of n observations, case i where $i = 1, 2, 3, , ..., n$ is considered randomly censored if $C_i < T_i$ and $C_i < C$, where C is the fixed type I censored time.

3.5.2 Left censoring, interval censoring and left truncation

A subject is said to be left censored if it is known that the event of interest occurred some time before the recorded follow-up period. For example, if a patient experience the first episode of peritonitis before the study begins, this patient is left censored for further analysis. A subject is said to be interval censored if it is known that the event of interest is located between two known time points, but the exact of failure is not known.

In a time to event data analysis, there exist a unique type of missing data called left truncation. A PD patients who enters the observation process after a given starting date is referred to as a staggered or delay entry and such kind of observations are said to be left truncated. Left truncation can potentially cause serious selection bias in survival analysis since it underestimates the risk of failure, however there are standard statistical techniques for handling such bias (Liu, 2012).

3.6 Continuous lifetime functions

In survival analysis, the random variable of interest is non-negative, usually denoted by T, called failure time, survival time, lifetime, time to event, etc. These terms are used interchangeably in this study. The dependent variable in survival analysis is divided into two parts, which are the time to event variable and event status variable, recording whether the patient is censored or not. The survival and hazard functions are the two quantitative terms in survival analysis for describing the distribution of time to event.

3.6.1 Survival function

The survival function is defined as the probability of not experiencing the event (surviving) beyond certain time period:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - \int_o^t f(u)du = 1 - F(t) \tag{3.1}$$

where $F(t)$, is the cumulative distribution function (cdf), defined as the probability that no event occurs over the time interval $(0, t)$. The survival functions have the following three properties:

- $S(t)$ is a non-negative function of t, that is, it approaches 0 as the time (t) increases.
- At time $t = 0$, the probability of surviving beyond t is 1, that is $S(t = 0) = P(T > 0) = 1$ since no event has occurred at the beginning of the study.
- At time $t = \infty$,the probability of surviving beyond time t is 0, that is $S(t = \infty) = 0$, meaning if the follow-up period increases without limit all the patients will get the peritonitis (event).

3.6.2 Hazard function

The hazard function at time t is defined as the instantaneous potential per unit time for the event to occur, given that an individual has survived up to some specified time:

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} \tag{3.2}$$

The hazard function is sometimes referred to as the conditional failure rate and it has the following properties: it is always non-negative and it has no upper bound.

3.6.3 Relationship between survival function and hazard function

There exist a straightforward relationship between the survival and the hazard functions:

$$\begin{aligned} h(t) &= lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} \\ &= lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T < t + \Delta t)}{P(T > t)} \\ &= lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{S(t) - S(t + \Delta t)}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

Conversely,

$$h(t) = \frac{f(t)}{S(t)} = \frac{-1}{S(t)} \frac{d}{dt} S(t) = -\frac{d}{dt} log S(t)$$

which imply that,

$$log(S(t)) = -\int_0^t h(u) du$$

and from this, we can exponentiate to get,

$$S(t) = exp(-\int_o^t h(u) du) = exp(-H(t))$$

where, $H(t)$ is the integration of all hazard rates over the time interval $(0, t)$, defined as the continuous cumulative hazard function at time t (Liu, 2012). Furthermore, the probability density function (pdf), $f(t)$ can be written in terms of the hazard function:

$$f(t) = h(t)S(t) = h(t)exp(-\int_o^t h(u) du = h(t)exp(-H(t)) \tag{3.3}$$

3.6.4 Mean survival

The expected time remaining at time t, also referred to as life expectancy at t can be computed using the basic function defined in section 3.6.3. A it represents the unit-based probability surviving at time t, $S(t)$ can be considered the intensity of expected life at time t. Taking a limit as $t \to \infty$ then $S(t) = 0$ then the expected life remaining at time 0 is given by

$$E(T_0) = E(T|t = 0) = \int_o^\infty S(u)du \tag{3.4}$$

Likewise, the expected life remaining at time t is given by

$$E(T_t) = E(T|T \geq t) = \frac{\int_t^\infty S(u)du}{S(t)} \tag{3.5}$$

where $S(t)$ represents exposure for the expected life remaining at time t. The distribution of the residual life time is given by,

$$F(u) = P(T - t \leq u|T \geq t) = \frac{F(t + u) - F(t)}{S(t)} \tag{3.6}$$

this leads to the mean survival time given by (Childers, 2015),

$$\mu = E(T) = \int_0^\infty tf(t) = \int_0^\infty S(t)dt \tag{3.7}$$

which is referred to as the average survival rate. Since the mean survival time normally is not estimable in the presence of right censoring, the restricted mean survival time can be used instead. With the Kaplan-Meier estimator μ an be estimated as

$$\hat{\mu} = E(T) = \int_0^\infty tf(t) = \int_0^\infty \hat{S}(t)dt \tag{3.8}$$

3.6.5 Median survival

In survival analysis, sometimes the median is more beneficial than the average (Childers, 2015). The p^{th} quantile of survival time is the smallest t_p such that:

$$S(t_p) \leq 1 - p; t_p = inf(u : S(u) \leq 1 - p). \tag{3.9}$$

However, for the continuous random variable T, the p^{th} quantile can be obtained from $S(t_p) = 1 - p$. It then follows that the 50^{th} percentile, $t_{0.5}$ of the distribution of T is given by $S(t_{0.5}) = 0.5$.

3.7 Non-parametric survival methods

The non-parametric methods are the techniques that does not make any assumptions on the form of the probability distribution, but rely completely on the empirical data. In survival analysis, the Kaplan-Meier and Nelson-Aalen estimators are the two well-known and related non-parametric techniques for analyzing the survival probability and the cumulative hazard function (Liu, 2012).

3.7.1 Formulation of the Kaplan-Meier and Nelson-Aalen estimators

Suppose that $t_1 < t_2 < ... < t_r$ are the ordered time to event, n_i is the number of individual at risk at time t_i, and d_i denote the number of events at time t_i. The conditional probability that an individual fail in the time interval $t_i - \Delta$ to t_i, given that the individual has survived up to time $t_i - \Delta$, is denoted by $\frac{d_i}{n_i}$, and the conditional probability that an individual survives past time $t_i - \Delta$, given that has survived up to time $t_i - \Delta$, is estimated by $\frac{n_i - d_i}{n_i}$, taking the limit as $\Delta \to 0$, $\frac{n_i - d_i}{n_i}$ gives the conditional probability of surviving past time t_i given that has survived up to time

t_i.

The probability of surviving past time t, for $t_k \leq t < t_{k+1}$, is given by

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(T > t \ \ and \ \ T > t_k) \\ &= P(T > t | T > t_k) P(T > t_k) \\ &= P(T > t | T > t_k) P(T > t_k | T > t_{k-1}) P(T > t_{k-1}) \\ &= P(T > t | T > t_k) P(T > t_k | T > t_{k-1}) \times \\ & P(T > t_{k-1} | T > t_{k-2}) \times ... \times P(T > t_1 | T > t_0) P(T > t_0) \\ &\approx \prod_{i=1}^{k} P(T > t_i | T > t_{i-1}) \end{aligned}$$

where $t_0 = 0$ and $t_{k+1} = \infty$.

The Kaplan-Meier estimator of the survival function at time t, for $t_k \leq t < t_{k+1}$, is given by

$$\widehat{S}(t) = \prod_{i=1}^{k} \frac{n_i - d_i}{n_i} \tag{3.10}$$

where, $\widehat{S}(t)$ is the Kaplan-Meier estimate for the probability of survival at time t. Censoring is not specified in this equation, in situation where censored time exist, the Kaplan-Meier estimator would be written as,

$$\widehat{S}(t) = \prod_{i=1}^{k} (\frac{n_i - d_i}{n_i})^{\delta_i} \tag{3.11}$$

where δ_i is the time status variable or censorship indicator, taking the value 0, $(\delta_i = 0)$, when time t_i is the censored survival time and the value 1, $(\delta_i = 1)$, when t_i is the actual survival time (Liu, 2012).

The Nelson-Aalen estimator can be derived from the Kaplan-Meier estimator, by converting the survival function to the cumulative hazard function, that is, from the

equation, $H(t) = -log[S(t)]$, it follows that,

$$\begin{aligned}\widehat{H}(t) &= -log[\widehat{S}(t)] \\ &= -log[\prod_{i=1}^{k}(\frac{n_i - d_i}{n_i})^{\delta_i}] \\ &= -\sum_{i=1}^{k}\delta_i log(\frac{n_i - d_i}{n_i}) \\ &= -\sum_{i=1}^{k}\delta_i log(1 - \frac{d_i}{n_i}) \\ &= -\sum_{i=1}^{k}\delta_i(-\frac{d_i}{n_i})\end{aligned}$$

by applying $log(1 + x) \approx x$, the Nelson-Aalen estimator is given by,

$$\widehat{H}(t) = \sum_{i=1}^{k}(\frac{\delta_i d_i}{n_i}) \tag{3.12}$$

When the censorship indicator is not considered, the Nelson-Aalen estimator is written as,

$$\widehat{H}(t) \approx \sum_{i=1}^{k}(\frac{d_i}{n_i}) = \prod_{i=1}^{k}[1 - \widehat{S}(t_i)] \tag{3.13}$$

where $\widehat{H}(t)$ is the estimate of the cumulative hazard function and $\widehat{S}(t_i)$ is the estimate of the probability of survival at time t_i

3.7.2 Greenwood's formula

For the large population sample, the Kaplan-Meier estimator approximates the mean of the survival probability, asymptotically normally distributed. Considering this property, the variance of the survival estimate can be derived for assessing the dispersion of the survival probability (Liu, 2012). In order to derive the variance of the survival estimate (Greenwood's formula), there are several transformation steps to be considered. The first step is take the natural logarithm of the Kaplan-

Meier survival function, that is,

$$log[\widehat{S}(t) = \sum_{i=1}^{k} log(\frac{n_i - d_i}{n_i})] = \sum_{i=1}^{k} log[\widehat{S}(t_i)]$$

where $\widehat{S}(t_i)$ is the conditional probability of survival in the time interval (t_{i-1}, t_i). The variance of $\widehat{S}(t_i)$, is given by

$$\widehat{V}[\widehat{S}(t_i)] = \frac{\widehat{S}(t_i)[1 - \widehat{S}(t_i)]}{n_i}$$

since $\widehat{S}(t_i)$ can be expressed as an estimate of the proportion (Liu, 2012). Applying the delta method, the variance of $log[\widehat{S}(t_i)]$ is approximated by,

$$\begin{aligned}\widehat{V}[log(\widehat{S}(t_i))] &\approx [\frac{1}{\widehat{S}(t_i)}]^2 \frac{\widehat{S}(t_i)[1 - \widehat{S}(t_i)]}{n_i} \\ &= \frac{1 - \widehat{S}(t_i)}{n_i \widehat{S}(t_i)} \\ &= \frac{n_i - n_i \widehat{S}(t_i)}{(n_i)^2 \widehat{S}(t_i)}\end{aligned}$$

when both numerator and denominator are multiplied by the common term n_i.

From $n_i \widehat{S}(t_i) = n_i - d_i$, the variance of $log[\widehat{S}(t_i)]$ can be written as,

$$\widehat{V}[log(\widehat{S(t_i)})] \approx \frac{d_i}{n_i(n_i - d_i)}$$

now, the variance of $log[\widehat{S}(t)]$ can be obtained by summing up the variances of all $log[\widehat{S}(t_i)]$, that is

$$\widehat{V}[log(\widehat{S}(t))] \approx \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)}$$

performing the re-transformation procedure using the delta method, the variance of the survival probability $\widehat{S}(t)$, known as the Greenwood's formula is given by

$$\widehat{V}[\widehat{S}(t)] \approx [\widehat{S}(t)]^2 \sum_{i=1}^{k} \frac{n_i}{n_i(n_i - d_i)}. \tag{3.14}$$

In a large sample the Kaplan-Meier estimator at time t, $S(t)$ is approximately normally distributed (Borgan, 1997). The $100(1-\alpha)\%$ confidence interval for $S(t)$ is given by,

$$\widehat{S}(t) \pm Z_{1-\frac{\alpha}{2}}\widehat{\sigma}(t) \tag{3.15}$$

where $\widehat{\sigma}(t)$ is the square root of the Greenwood's formula (Equation 3.12), known as the Greenwood's formula standard error. However, there is a serious shortcoming from using this formula for estimating the variance of the survival function. The probability of survival ranges between 0 and 1 and this formula can yield values greater than 1 or values less than 0. Given this concern, the confidence interval of $\hat{S}(t)$ needs to be estimated by some transformation approaches. The log-log transformation is the most popular transformation for estimating the confidence interval of $\hat{S}(t)$.

The logic of the log-log transformation is that te asymptotic normal distribution of $\hat{S}(t)$ should be first transformed to a continuous function with unrestricted bounds. The transformed survival function given by

$$y(t) = log[-log\hat{S}(t)] \tag{3.16}$$

Applying the delta method with respect to the Greenwood formula, the variance of $y(t)$ (equation 3.14) can be derived:

$$\hat{V}[y(t)] \approx \frac{[\hat{V}\hat{S}(t)]}{[\hat{S}(t)log\hat{S}(t)]} \tag{3.17}$$

Given Equation (3.15), the transformed log-log confidence interval for the survival function is given by

$$log[-log\hat{S}(t)] \pm \sqrt{\frac{[\hat{V}\hat{S}(t)]}{[\hat{S}(t)log\hat{S}(t)]}} \tag{3.18}$$

Where $z_{1-\frac{\alpha}{2}}$ is the z-score for the upper $\frac{\alpha}{2}$ percentile of the standard normal distribution.

3.7.3 Group comparison of survival functions

Kaplan-Meier estimator can be employed for comparing the survival functions by adding certain stratification factors. In comparing the survival functions between two or more population groups, an observed difference can either be a reflection of the sampling error or the outcome of an actual disparity (Liu, 2012). Given this information, it is important to conduct the significance tests for determining whether to the observed difference is true or not. When conducting the significance tests, the null and alternative hypothesis must be clearly stated, (H_0 :No statistically significant difference between groups vs H_1 :Statistically significant difference groups). The critical value and the $p - value$ can be used the conclude whether the null hypothesis, (H_0) should be rejected or fail to be rejected.

The normal and chi-squared distribution are the most widely used probability functions for the hypothesis testing. In survival analysis, there are many techniques for testing the significance difference between two or more survival functions or survival curves. The log-rank, Peto, Tarone and the Wilcoxon test are some of the methods. The log-rank is the most common used test and will be therefore discussed in this chapter.

Log-rank test

Instead of looking at the fixed time points, the log-rank test compare the whole survival function for association in different groups,

$\mathbf{H}_0 : S_1(t) = S_2(t) = ... = S_k(t)$

The true survival functions are not known for each group and therefore the non-

parametric test must be employed. Before deriving the non-parametric test statistic for this null hypothesis, the simple setting with $K = 2$ must be considered because it is more clear on which idea to be used.

$\mathbf{H}_0 : S_1(t) = S_2(t)$

When this null hypothesis of no association between two groups (G_1 and G_2) is true, the marginal totals in Table 3.1 should all be fixed and, consequently d_{1i} can be viewed as a hypergeometric distribution random variable with parameters n_i, n_{1i}, and d_i (Liu, 2012). The hypergeometric probability distribution of having d_{1i} in n_{1i}, given n_i, n_{1i}, and d_i, is defined by

$$P(Y_{1i} = d_{1i}) = \frac{\binom{d_i}{d_{1i}}\binom{n_i - d_i}{n_{1i} - d_{1i}}}{\binom{n_1}{n_{1i}}} \tag{3.19}$$

where Y_{1i} is the random variable for d_{1i}, d_{1i} is the number of failures from G_1, and n_{1i} is the number of individuals at risk in G_1 at time i. The hypergeometric distribution of d_{1i}, is well defined, with the mean

$$E(d_{1i}) = \frac{d_i n_{1i}}{n_i} \tag{3.20}$$

and the variance

$$Var(d_{1i}) = \frac{d_i(n_i - d_i)n_{1i}n_{21}}{n_i^2(n_i - 1)} \tag{3.21}$$

Table 3.1: Number of events and nonevents at t_i in two groups

Group	Event (Yes)	Event (No)	Total
G_1	d_{1i}	n_{1i} - d_{1i}	n_{1i}
G_2	d_{2i}	n_{2i} - d_{2i}	n_{2i}
$Total$	d_i	n_i - d_i	n_i

In 1959, Mantel and Haenszel proposed to sum the difference between the observed d_{1i} and the expected value of the observed $E(d_{1i})$, over all the observed

survival times. The sum is given by

$$D = \sum_{i=1}^{n} [d_{1i} - E(d_{1i})] \tag{3.22}$$

and the variance of the sum is expressed as

$$Var(D) = \sum_{i=1}^{n} var(d_{1i}) = \sum_{i=1}^{n} [\frac{d_i(n_i - d_i)n_{1i}n_{21}}{n_i^2(n_i - 1)}] \tag{3.23}$$

If the sample size at each failure time is sufficiently large, the sum, D, is approximately normally distributed and its standardised form is given by

$$Z = \frac{\sum_{i=1}^{n} [d_{1i} - E(d_{1i})]}{\sqrt{\sum_{i=1}^{n} [\frac{d_i(n_i - d_i)n_{1i}n_{21}}{n_i^2(n_i - 1)}]}} = \frac{D}{\sqrt{Var(D)}} \sim N(0, 1) \tag{3.24}$$

which is the standard Z-test statistic.

Statistically, the square of the Z-test statistic gives the chi-square distribution, and the efficient test statistic based on the chi-square distribution is called the log-rank test statistic, generally given by

$$Q_{logrank} = \frac{D^2}{Var(D)} \tag{3.25}$$

which is, under $\mathbf{H}_0$, approximately distributed as chi-square with one degrees of freedom for two groups (Liu, 2012).

Suppose that $K > 2$, i.e there are more than two different groups to be compared. The null hypothesis to be tested is

$\mathbf{H}_0 : S_1(t) = S_2(t) = ... = S_k(t)$

i.e No significant difference between all K survival functions. This hypothesis is tested by generalising the test of two groups. where d_{1i} is the number of failures from G_1, and n_{1i} is the number of individuals at risk in G_1 at time i.

Under $\mathbf{H}_0$, the vector $\mathbf{O}_i = [d_1, d_2, ..., d_{k-1}]$ of the observed number of events in

Table 3.2: Number of events and nonevents at t_i in K groups

Group	Event (Yes)	Event (No)	Total
G_1	d_{1i}	n_{1i} - d_{1i}	n_{1i}
G_2	d_{2i}	n_{2i} - d_{2i}	n_{2i}
-	-	-	-
G_k	d_{ki}	n_{ki} - d_{ki}	n_{ki}
$Total$	d_i	n_i - d_i	n_i

group 1 to $(K-1)$ at failure time t_i, follows a multivariate hypergeometric distribution with mean vector $\mathbf{E}_i$ and variance-covariance matrix $\mathbf{V}_i$, where

$$\mathbf{E}_i = [\frac{d_i n_{1i}}{n_i}, \frac{d_i n_{2i}}{n_i}, ..., \frac{d_i n_{(k-1)i}}{n_i}]' \tag{3.26}$$

The K^{th} diagonal elements in the variance-covariance matrix $(\mathbf{V}_i)$ is given by

$$\mathbf{V}_{kki} = \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)} \tag{3.27}$$

and the K^{th} off-diagonal element is given by

$$\mathbf{V}_{kli} = \frac{n_{ki}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)} \tag{3.28}$$

for $k \neq l$

Generalising the log-rank test statistic for two group, result on

$$Q_{logrank} = (\mathbf{O} - \mathbf{E})'\mathbf{V}^{-1}(\mathbf{O} - \mathbf{E}) \sim X^2(K-1) \tag{3.29}$$

where $\mathbf{O} = \sum_{i=1}^{n} \mathbf{O}_i$, $\mathbf{E} = \sum_{i=1}^{n} \mathbf{E}_i$, and $\mathbf{V} = \sum_{i=1}^{n} \mathbf{V}_i$

3.8 Survival distributions functions

The distribution of event time T usually follows a predictable pattern. In this case parametric models can be developed for describing the survival processes of the

time to event. In survival analysis, these distributions are employed to perform the test survival differences between two or more groups. These distributions can reliably predict time to event well after the period during which events occurs (Stevenson and EpiCentre, 2009). Their parameters are often estimated by using the appropriate modification of the maximum likelihood. The popular distributions in survival analysis includes the exponential, Weibull, Gamma, Log-normal and Gompertz distribution.

Exponential distribution

The exponential distribution is the simplest function among the families of the parametric time distributions because its specification is based on a single parameter, λ (Liu, 2012). A characteristic of the exponential distribution is that the instantaneous hazard function does not vary over time t, (the hazard function is constant). The survival function at time t with the exponential distribution is given by,

$$S(t;\lambda) = exp(-\lambda t) \tag{3.30}$$

for $t > 0$, where λ, is the constant rate of change throughout the time interval $(0, \infty)$.

The hazard function in the exponential distribution can be derived as follows,

$$h(t) = -log\frac{d}{dt}S(t;\lambda) = \lambda \tag{3.31}$$

for $t > 0$. This constant rate, λ determines the scale of the hazard function and sometimes it is referred to as the scale parameter. Given the hazard and survival functions, the p.d.f can be obtained as follows:

$$f(t;\lambda) = h(t;\lambda)S(t;\lambda) = \lambda e^{-\lambda t} \tag{3.32}$$

for $t > 0$. The expected life of the exponential time distribution can be derived as

fallows,

$$E(T) = \int_0^\infty t\lambda e^{-\lambda t} = \frac{1}{\lambda} \tag{3.33}$$

The median value of the exponential survival function, denoted by t_p, can be obtained from λ, by using the condition that $S(t_p) = 0.5$

$S(t_p) = 0.5 \Rightarrow \lambda e^{-\lambda t} = 0.5 \Rightarrow -\lambda t_p = -log(2)$, therefore

$$t_p = \frac{1}{\lambda} log(2) \tag{3.34}$$

The exponential distribution posses the essential property called lack of memory. The probability suggest that the probability of surviving another t time units does not depend on how long one have lived.

$$P(T > t) = P(T > t + t_0 | T > t_0) \tag{3.35}$$

for any $t_0 > 0$.

Weibull distribution

The Weibull distribution is the most widely employed parametric function in survival analysis because of its flexibility and simplicity. It is described by the two parameters, scale parameter λ and a shape parameter α. The characteristics of the Weibull distribution is that the instantaneous hazard function is monotonically decreasing when $\alpha < 1$, monotonically increasing when $\alpha > 1$, and equivalent to the exponential distribution when $\alpha = 1$.

The survival function at time t with the Weibull distribution is given by,

$$S(t; \lambda, \alpha) = exp[-(\lambda t)^\alpha] \tag{3.36}$$

for $t > 0$. The condition that, $h(t; \lambda, \alpha) = -\frac{d}{dt} log[S(t; \lambda, \alpha)]$ lead to the following

equation:

$$h(t;\lambda,\alpha) = \alpha\lambda^{\alpha}t^{\alpha-1} \tag{3.37}$$

$t > 0$, is called the hazard function in the Weibull distribution. The specification of the hazard and survival functions lead to the p.d.f given by,

$$f(t;\lambda,\alpha) = h(t;\lambda,\alpha)S(t;\lambda,\alpha) = \alpha\lambda^{\alpha}t^{\alpha-1}exp[-(\lambda t)^{\alpha}] \tag{3.38}$$

for $t > 0$. The expected survival denoted by, $E(T;\lambda,\alpha)$ is given

$$E(T;\lambda,\alpha) = \frac{1}{\lambda}\Gamma(1+\frac{1}{\alpha}) \tag{3.39}$$

The median survival time of the Weibull distribution obtained from $S(t_p) = 0.5$ is given by:

$$t_p = [\frac{log(2)}{\lambda}]^{\frac{1}{\alpha}} \tag{3.40}$$

Gamma distribution

The gamma distribution can be described as the two-parameter family of continuous distributions. It has a shape parameter, denoted by $\beta > 0$ and a scale parameter defined by $\lambda > 0$. The gamma distribution arises naturally (there are real-life for which an associated survival distribution is approximately gamma) as well as analtically (simple function of random variable have a gamma distribution). The p.d.f is given by:

$$f(t;\lambda,\beta) = \frac{\lambda^{\beta}t^{\beta-1}e^{-\lambda t}}{\Gamma(\beta)} \tag{3.41}$$

for $t > 0$, $\lambda > 0$, and $\beta > 0$, where the gamma function $\Gamma(\beta)$, is defined by $\Gamma(\beta) = \int_0^{\infty} t^{\beta-1}e^{-t}dt$. The survival function at time t with the gamma distribution is defined by:

$$S(t;\lambda,\beta) = 1 - \frac{1}{\Gamma(\beta)}\int_0^{\lambda t} u^{\beta-1}e^{-u}du \tag{3.42}$$

for $t > 0$.

Log-normal distribution

The log-normal distribution is another widely used parametric function in survival analysis. The lifetime variable T has a log-normal distribution if $log(T)$ has a normal distribution and this serves as its fundamental characteristic. This characteristic makes the specification of the log-normal distribution uncomplicated as compared to the Weibull and the gamma distribution (Liu, 2012).

The density function of the log-normal distribution can be expressed as,

$$f(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} exp - \frac{1}{2\sigma^2}(logt-\mu)^2 = \frac{1}{t}\phi(\frac{logt-\mu}{\sigma}) \tag{3.43}$$

for $t > 0$, where $\phi(.)$ is the p.d.f of the normal distribution. The survival function is given by:

$$S(t;\mu,\sigma) = 1 - \varphi(\frac{logt}{\sigma}) \tag{3.44}$$

for $t > 0$, where $\varphi(.)$ represent the cumulative normal distribution, also known as the probit model. The hazard function is given by:

$$h(t;\mu,\sigma) = \frac{1}{t\sigma}\phi(\frac{logt}{\sigma})/\varphi(\frac{-logt}{\sigma}) \tag{3.45}$$

for $t > 0$.

Gompertz distribution

The lifetime variable T has the Gompertz distribution if the density function is given by:

$$f(t;\alpha,\beta) = \beta e^{\alpha t} exp[\frac{\beta}{\alpha}(1-e^{\alpha t})] \tag{3.46}$$

for $t > 0$, where β is called the age-independent hazard rate coefficient and α is referred to as the age-independent mortality rate coefficient (Liu, 2012). The Gompertz survival function at time t can be derived from integrating the p.d.f, given

by:

$$S(t;\alpha,\beta) = exp[\int_0^t h(u;\alpha,\beta)du] = exp[\frac{\beta}{\alpha}(1-e^{\alpha t})] \tag{3.47}$$

Given the survival and the density functions, the hazard function can be derived as follows:

$$h(t;\alpha,\beta) = \frac{f(t;\alpha,\beta)}{S(t;\alpha,\beta)} = \beta e^{\alpha t} \tag{3.48}$$

for $t > 0$. The log transformation of the Gompertz hazard function is linearly associated with time t, ie,

$$log[h(t;\alpha,\beta)] = log\beta + \alpha t \tag{3.49}$$

3.9 Parametric regression models

In survival analysis, regression models are used for assessing the association between the outcome lifetime variable and one or more independent variables, with one or more variables serving as controls (Liu, 2012). In parametric regression, the dependent variable is the event time T, which maybe censored and it is assumed to follow a known probability distribution whose parameter(s) may depend on the covariates, denoted by **X**. The effects of this covariates can be modelled by using either survival time or hazard rate at time t as a function of a parameter vector $\boldsymbol{\theta}$, (Liu, 2012).

Parametric regression modeling can be viewed in different perspectives, however, the two most popular perspectives are the parametric hazard rate model and the log transformation (Liu, 2012). In parametric hazard rate, the proportional hazard rate is most common and the effects of covariates are assumed to be multiplicative. In the log transformation of event times, the accelerated failure time model (AFT model) is widely used and it is assumed to be linearly associated with the covariates.

3.9.1 Proportional hazard regression model

The proportional hazard regression, models the hazard rate,(the number of new cases at-risk per unit time). The hazard function at time t for an individual with covariate **X** is assumed to be,

$$h(t|\mathbf{X}) = h_o(t)exp[\boldsymbol{\beta}'\mathbf{X}] \tag{3.50}$$

where $h_0(t)$ denotes a known baseline hazard function for a continuous time variable T, that describes the risk for individuals when all covariates $X_1, X_2, X_3, ..., X_p$ equal to 0., that is, $h(t|\mathbf{X} = 0) = h_0(t)$

These models are referred to as the proportional hazard models because the ratio for the two different individuals is constant, that is,

$$\frac{h_0(t|\mathbf{X}^{'})}{h_0(t|\mathbf{X}^{'})} = \frac{h_0(t)e^{\boldsymbol{\beta}'\mathbf{X}}}{h_0(t)e^{\boldsymbol{\beta}'\mathbf{X}^{'}}} = e^{\boldsymbol{\beta}'(\mathbf{X}-\mathbf{X}^{'})} \tag{3.51}$$

Using the relationship between the hazard and survival functions, the survival function for time T given **X** from baseline hazard function is derived as follows,

$$\begin{aligned} S(t|\mathbf{X}) &= exp[-\int_0^t h_0 h(u)exp(\boldsymbol{\beta}'\mathbf{X})du] \\ &= exp[-exp(\boldsymbol{\beta}'\mathbf{X})\int_0^t h_0 h(u)du] \\ &= exp[-exp(\boldsymbol{\beta}'\mathbf{X})H_0(t)] \\ &= exp[-H_o(t)]^{\boldsymbol{\beta}'\mathbf{X}} \\ &= [S_0(t)]^{exp(\boldsymbol{\beta}'\mathbf{X})} \end{aligned} \tag{3.52}$$

where $S_0(t)$ is the baseline survival function and $H_0(t)$ is defined as the continuous cumulative baseline hazard function at time t.

Given the survival and hazard functions, the p.d.f of time variable T given **X** is

defined as,

$$f(t|\mathbf{X}) = h_o(t)exp(\boldsymbol{\beta}'\mathbf{X}) = exp[-exp(\boldsymbol{\beta}'\mathbf{X}) \int_0^t h_0 h(u)du] \tag{3.53}$$

Different kinds of the proportional hazard models may be derived by making different assumptions on the baseline survival function or the baseline hazard function. for example if the baseline hazard is constant, $(h_0(t) = \lambda)$ throughout the follow-up interval, the exponential regression model given by

$$h(t|\mathbf{X} : T \sim Exp(\lambda)) = \lambda exp(\boldsymbol{\beta}'\mathbf{X}). \tag{3.54}$$

is obtained.

Which indicates that the exponential regression model on the hazard rate is just the product of a constant baseline rate and a multiplicative form $exp(\boldsymbol{\beta}'\mathbf{X})$, representing the effect of the covariates vector **X**, (Liu, 2012).

When the observed hazard function varies monotonically over time, the Weibull regression model given by

$$h(t|\mathbf{X} : T \sim Wei(\alpha, \lambda)) = \alpha\lambda(\lambda t)^{\alpha-1} exp(\boldsymbol{\beta}'\mathbf{X}) \tag{3.55}$$

where $h_0(t) = \alpha\lambda(\lambda t)^{\alpha-1}$ is the Weibull baseline rate.

3.9.2 Accelerated failure time regression model

Parametric regression models in survival analysis can be created on the log time T over covariates, from which a different set of parameters needs to be specified (Liu, 2012). This type of regression model is referred to as the AFT regression model. The general form of the AFT is defined as,

$$Y = log(T) = \mu + \mathbf{X}'\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon} \tag{3.56}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients on $logT$, μ is the intercept parameter, σ is the scale parameter and $\boldsymbol{\varepsilon}$ is the error term that follows a particular parametric distribution of time T with survival function $S(t)$. The AFT model show the log-linear association between time T and the covariate vector $\mathbf{X}$. With the location term $(\mathbf{X}^{'}\boldsymbol{\beta})$ and scale parameter σ, the baseline parametric distribution of survival time can be conveniently modelled by a term of additive random disturbances $(logT_0)$ (Liu, 2012). The survival function for the i^{th} individual, using the equation $logT \geq logt$, is given by

$$\begin{aligned} S_i(t) &= P[\mu + \mathbf{X}_i^{'}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i} \geq logt] \\ &= P[\boldsymbol{\varepsilon_i} \geq \frac{logt - \mu - \mathbf{X}_i^{'}\boldsymbol{\beta}}{\sigma}] \end{aligned} \tag{3.57}$$

where $\boldsymbol{\varepsilon_i}$, is the component of the error vector $\boldsymbol{\varepsilon}$, $S(t) = P(\boldsymbol{\varepsilon_i} \geq t)$, $F(t) = P(\boldsymbol{\varepsilon_i} < t)$ and $f(t) = \frac{d}{dt}F(t)$. Because the explanatory variable with coefficients, $(\mathbf{X}^{'}\boldsymbol{\beta})$ is independent of the distribution parameter $\boldsymbol{\varepsilon_i}$, the survival function $S(t)$ with respect to $logT$ can be modelled by specifying a random component and a fixed component, given by

$$S(t|\mathbf{X}) = S_0(\frac{logt - \mu - \mathbf{X}^{'}\boldsymbol{\beta}}{\sigma}) \tag{3.58}$$

for $-\infty < logt < \infty$, where S_0 is the survival function of the distribution of $\boldsymbol{\varepsilon}$ and $\mathbf{X}^{i}\boldsymbol{\beta}$ defines the location of time T, called an accelerated factor.

The cumulative hazard function can be expressed using $H(t) = -logS(t)$, that is,

$$H(t) = logS_0(\frac{logt - \mu - \mathbf{X}^{'}\boldsymbol{\beta}}{\sigma}) = H_0(\frac{logt - \mu - \mathbf{X}^{'}\boldsymbol{\beta}}{\sigma}) \tag{3.59}$$

where H_0 is the cumulative hazard of $\boldsymbol{\varepsilon}$.

Using $h(t) = \frac{d}{dt}[-logS(t)]$, the hazard function as defined as

$$h(t|x) = \frac{d}{dt}H_0(\frac{logt - \mu - \mathbf{X}^{'}\boldsymbol{\beta}}{\sigma}) = \frac{1}{\sigma t}h_0(\frac{logt - \mu - \mathbf{X}^{'}\boldsymbol{\beta}}{\sigma}) \tag{3.60}$$

where h_0 is the baseline hazard function for the distribution $\boldsymbol{\varepsilon_i}$, also independent of

X.

Because of the simplicity and flexibility of the log-linear model, the survival function can be applied to formulate a large number of families of parametric distributions in survival analysis. The effect of covariate vector **X** on time T changes location, but not the shape of the distribution of T, thus the parametric regression models that can be formulated by the survival functions are referred to as the AFT regression models, (Liu, 2012). The advantage of using the AFT model is that it covers a wide range of the survival time distribution. The AFT regression model can be formulated with respect to the random variable T, rather than $logT$, that is,

$$T = exp(\mu + \mathbf{X}^{'}\boldsymbol{\beta})exp(\sigma\boldsymbol{\varepsilon}) = exp(\mu + \mathbf{X}^{'}\boldsymbol{\beta})\mathbf{E} \tag{3.61}$$

where $\mathbf{E} = exp(\sigma\boldsymbol{\varepsilon}) > 0$ has the hazard function $h_0(\mathbf{e})$ and is independent of $\boldsymbol{\beta}$. Because $exp(\sigma\boldsymbol{\varepsilon}$ is positive valued, T is restricted in the range $(0, \infty)$. Using $T \geq t$, the survival function for the i^{th} individual is expressed as,

$$S_i(t) = P(T_i \geq t) = P[exp(\mu + \mathbf{X}^{'}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}_i) \geq t] \tag{3.62}$$

Since the term $\mathbf{X}^{'}\boldsymbol{\beta}$ is independent of the disturbance parameter, the survival function $S(t)$ can be written as

$$S_i(t) = S_0[t\ exp(\mathbf{X}^{'}\boldsymbol{\beta})] \tag{3.63}$$

where S_0 is a fully specified survival function, defined as $S_0(t) = exp(\mu + \mathbf{X}^{'}\boldsymbol{\beta} \geq t)$

Using the intimate relationship of the hazard and the survival function, the hazard function for T can be written as

$$h(t|x) = h_0[texp(-\mathbf{X}^{'}\boldsymbol{\beta})]exp(-\mathbf{X}^{'}\boldsymbol{\beta}) \tag{3.64}$$

Using $S(t) = exp(\int_0^t h(u)du)$, we get the survival function in terms of the AFT haz-

ard function:

$$S(t|x) = exp- \int_0^t h_0[uexp(-\mathbf{X}^{'}\boldsymbol{\beta})]exp(-\mathbf{X}^{'}\boldsymbol{\beta})du = expH_o[texp(\mathbf{X}^{'}\boldsymbol{\beta})] \tag{3.65}$$

The density function in the formation of AFT perspective is defined as,

$$f(t|x) = h_0[t\ exp(\mathbf{X}^{'}\boldsymbol{\beta})]exp(-\mathbf{X}^{'}\boldsymbol{\beta})expH_o[texp(\mathbf{X}^{'}\boldsymbol{\beta})] \tag{3.66}$$

which is obtained using $f(t) = h(t)s(t)$.

In the AFT regression models, the effect of covariates determine the time scale in such a way that if $exp(-\mathbf{X}^{'}\boldsymbol{\beta}) > 1$, the survival process accelerates and if $exp(-\mathbf{X}^{'}\boldsymbol{\beta}) < 1$, the survival process decelerates.

3.9.3 Cox proportional hazard regression model

It is difficult in practice to ascertain the correct underlying parametric distribution for survival times. Reseachers are more interested in how covariates influences the risk of an event occurrence than in the shape of a specific failure time distribution (Liu, 2012). Because of these concerns, it is useful to create a regression model that provides a valid estimates of the covariates effects on the hazard function while avoiding the specification of an underlying distribution function. The Cox proportional hazard regression model, introduced by Cox (1972), derives the efficient estimates of the covariate effects using the proportional hazard assumption while leaving the baseline hazard unspecified (Liu, 2012).

Cox semi-parametric hazard model

The Cox model has become the most popular and widely used regression model in survival analysis (Childers, 2015). The Cox model uses the maximum likelihood algorithm for the partial likelihood function, with the estimating procedure known

as a partial likelihood (Liu, 2012). In the Cox model, the conditional hazard of an individual, given the covariate values $X_1, X_2, X_3, ..., X_p$ is expressed as

$$h(t|\mathbf{X}) = h_0(t)exp(\boldsymbol{\beta}'\mathbf{X}) \tag{3.67}$$

where the multiplicative term $exp(\boldsymbol{\beta}'\mathbf{X})$ specifies the effect of covariates and the term $h_0(t)$ represent an arbitrary and unspecified baseline hazard function for continuous time variable T. The coefficient vector $\boldsymbol{\beta}$ provides a set of covariate effects on the hazard rate, with the same lengh as **X** and exponentiating a specific regression coefficient generates the hazard ratio (HR) of covariate. If for example, the covariate X_m is dichotomous variable with $X_m = 1$ and $X_m = 0$, and the other covariates takes the value 0, the hazard of covariate X_m is given by

$$HR_m = \frac{h_o(t)exp(X_{m1}\widehat{\beta}_m)}{h_o(t)exp(X_{m0}\widehat{\beta}_m)} = exp[(X_{m1} - X_{m0})\widehat{\beta}_m] = exp(\widehat{\beta}_m) \tag{3.68}$$

where $\widehat{\beta}_m$ is the estimate of the regression coefficient for covariate X_m. This definition of the hazard ratio(relative risk), independent of $h_0(t)$, holds when other covariates are not zero because additional terms appearing in both the numerator and denominator would cancel out (Liu, 2012).

For continuous covariate, the hazard ratio displays the extent to which the risk increases $(HR > 1)$ or decreases $(HR < 1)$ with 1-unit increase in the value of the covariate. In continuous setup the hazard ratio can also calculated to reflect the proportional change in the hazard rate with w-unit increase in X_m, given by

$$\begin{aligned} HR_m &= \frac{h_o(t)exp[(X_{m0} + w)\widehat{\beta}_m]}{h_o(t)exp(X_{m0}\widehat{\beta}_m)} \\ &= exp[(X_{m0} + w - X_{m0})\widehat{\beta}_m] \\ &= exp(w\widehat{\beta}_m) \\ &= exp(\widehat{\beta}_m)^w \end{aligned} \tag{3.69}$$

The measure $exp(w\widehat{\beta}_m)$ is referred to as the w-unit hazard ratio. This reflects the

multiplicative change in the hazard rate with a w-unit increase in the covariate X_m.

The Cox model can be expressed i n terms of the log-linear model (AFT), with covariates assumed to be linearly associated with the $logh(t)$ function. The AFT of the Cox model is given

$$log[h(t|\mathbf{X})] = log[h_0(t)] + \mathbf{X}\boldsymbol{\beta} = \alpha + \mathbf{X}\boldsymbol{\beta} \tag{3.70}$$

where α is the log of the baseline hazard function specified as the intercept factor in the parametric hazard regression model. This intercept is unspecified in the Cox model because $h_0(t)$ is unspecified.

3.10 Recurrent Survival analysis

Survival analysis of recurrent episodes considers a situation where an individual may experience more than one episode over the follow-up period. There are many techniques that can be utilised to handle analysis of this nature. However, this study will utilise the stratified, gap-time, marginal and the counting process approach. Irrespective of which approach is being utilised, the variance of the estimated partial likelihood regression coefficients should be adjusted for the possible correlation among recurrent episodes on the same individual. The sandwich robust variance estimator is the most popular and widely used estimator for adjusting the variances of the partial likelihood estimated regression coefficients.

3.10.1 Parametric regression models

The parametric regression models for modelling the time to recurrent episodes works the same way as the ordinary parametric regression models in the sense that they also assess the association between the outcome lifetime variable and one or more covariates.

Stratified model

In the stratified hazard regression model the proportional hazard regression model is assumed to be conditional for each individual. This model is assumed to be conditional simply because it considers that it is not possible for an individual to experience the k^{th} episode before experiencing the $(k-1)^{th}$ episode (i.e. an individual can not be at the risk set for subsequent episode without having experienced the preceding episode). Furthermore, each episode is considered as a separate process irrespective of whether it is coming from the same individual or not. The stratified hazard regression model for the i^{th} individual in the j^{th} stratum is of the form

$$h_{ij}(t|\mathbf{X}_{ij}) = h_{0j}(t)exp(\boldsymbol{\beta}'\mathbf{X}_{ij}) \tag{3.71}$$

where $h_{0j}(t)$ is the arbitrary and unspecified baseline hazard function for the continuous time variable T from stratum j, t represent the time interval (time-start to time-stop) for each follow-up period, $\mathbf{X}_{ij}$ is the vector of the i^{th} individual's covariate in the j^{th} stratum and $\boldsymbol{\beta}$ donate the set of the partial likelihood estimated regression coefficients.

This model is very useful when modelling the total time period of the recurrent episode process. That is, stratified model is more appropriate when modelling time-start to time-stop interval of the recurrent episode process. The variable stratum is used in this model to ensure that it is not possible for an individual to be in the risk set for subsequent episodes without having experienced the preceding episode.

Gap-time model

The gap-time hazard regression model is the most often utilised and more appropriate technique to employ when studying recurrent episodes rate as a function of time since the last episode (Duchateau et al., 2003). The technique operates

conditionally like the stratified process, however they differ only on how the time intervals are structured. In gap-time model, an individual moves to the k^{th} stratum immediately after the $(k-1)^{th}$ recurrence time and remains there until he or she experience the k^{th} episode or until the individual is censored. For instance, if an individual has one episode, then there will be two observations. An individual will move from the first stratum to the second stratum after experiencing the episode and remains there until the study period end (that is, until the individual is censored). In general, an individual with k episodes contributes $k+1$ observations. The gap-time hazard regression model for the i^{th} individual in the j^{th} stratum is expressed as

$$h_{ij}(t|\mathbf{X}_{ij}) = h_{0j}(t)exp(\boldsymbol{\beta}'\mathbf{X}_{ij}) \tag{3.72}$$

where $h_{0j}(t)$ represent an arbitrary and unspecified baseline hazard function for the continuous time variable T from stratum j. In the gap-time model t denote the gap between the time-start and time-stop for each follow-up period, $\mathbf{X}_{ij}$ is the vector of the i^{th} individual's vector in the j^{th} stratum and $\boldsymbol{\beta}$ represents the vector of the partial likelihood estimated regression coefficients.

The gap-time model is very useful when the researcher is interested in modelling the gap time between each time interval of the recurring episode rather than modelling the total time follow-up period of the recurrent episode process. Just as in the stratified model, the variable stratum is used in this model to ensure that it is not possible for an individual to be in the risk set for subsequent episodes without having experienced the preceding episode.

These models (gap-time and stratified) were established in 1981 and they are sometimes referred to as conditional Cox-type models, (Prentice et al., 1981). These models allow the shape of the hazard function to depend on the number of previous episodes and perhaps on the characteristics of the number of episodes an individual experiences and the covariate of an individual .

Marginal model

In the marginal model a proportional hazard regression model is assumed to be marginal for each individual, that is, each episode from the same individual is considered as a separate process. Furthermore, each individual is considered to be at the risk set for subsequent episodes, irrespective of the number of episodes each has actually experienced. In this model all individuals in the study contribute the follow-up times to all possible recurrent episodes. Furthermore, the marginal model considers each episode from the same individual separately and models all the obtainable information for the particular episode. The marginal hazard regression model for the i^{th} individual in the j^{th} stratum is of the form

$$h_{ij}(t|\mathbf{X}_{ij}) = h_{0j}(t)exp(\boldsymbol{\beta}'\mathbf{X}_{ij}) \tag{3.73}$$

where $h_{0j}(t)$ is the arbitrary and unspecified baseline hazard function for the continuous time variable T from stratum j, t represent the time at which the episode occurred or the time at which the individual got censored for each follow-up period. $\mathbf{X}_{ij}$ is the vector of the i^{th} individual's covariate in the j^{th} stratum and $\boldsymbol{\beta}$ denote the set of the partial likelihood estimated regression coefficients.

This model is very useful when modelling the exact time an individual experienced the episode or the exact time the individual got censored. That is, the marginal model is more appropriate when modelling the time-stop of the recurrent episode process. Just as in the stratified and gap-time model, the variable stratum is used in this model also to ensure that it is not possible for an individual to be in the risk set for subsequent episodes without having experienced the preceding episode.

Counting process model

In the counting process model each episode is assumed to be independent, regardless of whether it is coming from the same individual or not. Moreover, an

individual contributes to the risk set for an episode as long as the individual is still under the follow-up period during the time the episode occurred. The counting process model only differ from the other recurrent modelling technique because it does not take into account the order of the episodes. That is, the individual remains at the risk set for subsequent episode as long as they are still under the follow-up process at the time the episode occurs. This implies that the occurrence of the second episode does not depend on the occurrence of the preceding episode and individuals could be at risk set for subsequent episode without having experienced the preceding episodes. The counting process model for the i^{th} individual is expressed as

$$h_i(t|\mathbf{X}_i) = h_0(t)exp(\boldsymbol{\beta}'\mathbf{X}_i) \tag{3.74}$$

where $h_0(t)$ is the unspecified arbitrary baseline hazard function for the continuous time variable T, t represent the time-start and time-stop for each time interval of follow-up period. $\mathbf{X}_i$ is the vector of the i^{th} individual's covariate and $\boldsymbol{\beta}$ denote the set of the partial likelihood estimated regression coefficients.

The model that is used in the counting process approach is the same as the standard Cox proportional hazard model. However, in the counting process approach, an individual may experience more than one episode and these episodes are assumed to be independent from each other in a way that they are treated as if they are coming from different individuals.

The difference between the counting process model and the other three recurrent survival models is that the counting process is not conditional. That is, the counting process ignores the order at which the episodes occurred (i.e. individuals could be at the risk set for subsequent episodes without having experienced the preceding episodes). Furthermore, the counting process and stratified model uses the same time interval of the subsequent episode.

3.11 Model selection

In regression, it can be difficult to find a good model, especially in cases where there are many covariates. The criterion of identifying interesting covariates should be employed in such cases. The approach that will be utilised in model building for this study is the Akaike information criterion (AIC) established by Akaike in 1973 (Kleinbaum and Klein, 2006). The AIC investigate the likelihood and the number of parameters included in the model. It seek to balance the need of the model which fits the data well to that of having simple model. The AIC statistics is defined by

$$AIC = -2logL + \alpha q \tag{3.75}$$

where q is number of the unknown regression parameters in the model, α is any predetermined constant and L is the likelihood function. The AIC depends on the number of variables added in the model, in a way that, it will decrease as variables are being added in the model and increase when the number of added variables are unnecessary (Kleinbaum and Klein, 2006). The model with the smallest AIC value is more preferable.

The other criterion used to detect a good or a better fitting model is the Schwarz Bayesian Criterion (SBC). The SBC was derived from the Bayesian modification of the AIC criterion by Schwarz in 1978. This criterion was established for model selection. It is a function of the number of observations in the study, the sum of square error (SSE) and the number of independent covariates, $\alpha \leq p + 1$, where α considers the intercept also. The mathematical equation for this criterion is expressed as

$$SBC = nln(\frac{SSE}{n}) + \alpha lnn \tag{3.76}$$

The SBC criterion works like the AIC in a way that the model with the smallest value is considered to be the good or the best model to fit the available data set.

3.12 Partial likelihood

Suppose that the study is constructed on a sample of size n containing the triplets $(\mathbf{T}_j, \boldsymbol{\delta}_j, \mathbf{X}_j)$, $j = 1, 2, ..., n$, where $\mathbf{X}_i$ is a vector representing the set of covariates, $\mathbf{T}_i$ is a vector of the failure time random variable and $\boldsymbol{\delta}_i$ is a set censoring variable. The partial likelihood is derived under the following assumptions: There are ties between the failure times in the data and that given $\mathbf{X}_j$, censoring is non-informative (i.e. the failure time and censoring time for the j^{th} individual are independent). Let that $t_1 < t_2 < ... < t_D$ denote the D ordered distinct failure times and $X_{(i)k}$ denote the k^{th} covariate associated with an individual who experienced the event of interest at time t_i (Klein and Moeschberger, 2003). Suppose that $\Re(t_i)$ represent the set of all individuals who are at risk for failure at the time just before time t_i.

The probability that an individual experience an episode at time t_i with covariate $\mathbf{X}_j$, given that one of the individuals in the risk set experiences the episode at the very same time t_i, is represented by

$$P[individual\ i\ with\ covariate\ \mathbf{X}_j\ fails\ at\ time\ t_i | individual\ j\ from\ \Re(t_i)\ fails\ at\ t_i] \tag{3.77}$$

The conditional probability transforms a continuous hazard function and Equation (3.75) becomes

$$\begin{aligned} \frac{hazard\ rate\ at\ t_i\ for\ individual\ i\ with\ covariate\ \mathbf{X}_j}{\sum_{j\in\Re(t_i)} hazard\ rate\ at\ t_i\ for\ individual\ j} &= \frac{h[t_i|\mathbf{X}_{(i)}]}{\sum_{j\in\Re(t_i)} h[t_i|\mathbf{X}_j]} \\ &= \frac{h_0(t_i)exp[\boldsymbol{\beta}\mathbf{X}_{(i)}]}{\sum_{j\in\Re(t_i)h_0(t_i)exp[\boldsymbol{\beta}\mathbf{X}_j]} \\ &= \frac{exp[\boldsymbol{\beta}\mathbf{X}_{(i)}]}{\sum_{j\in\Re(t_i)exp[\boldsymbol{\beta}\mathbf{X}_j]} \end{aligned} \tag{3.78}$$

The partial likelihood formed by multiplying the conditional probabilities in equation (3.68) over all failures is based on the hazard function (Klein and Moeschberger,

2003) and is given by

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{D} \frac{exp[\boldsymbol{\beta}\mathbf{X}_{(i)}]}{\sum_{j\in\Re(t_i)exp[\boldsymbol{\beta}\mathbf{X}_j]} \\ &= \prod_{i=1}^{D} \frac{exp[\sum_{k=1}^{p}\beta_k X_{(i)k}]}{\sum_{j\in\Re(t_i)exp[\sum_{k=1}^{p}\beta_k X_{jk}]} \end{aligned} \tag{3.79}$$

when the censoring variable is considered, equation (3.69) can be expressed as follows

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{D} \{\frac{exp[\boldsymbol{\beta}\mathbf{X}_{(i)}]}{\sum_{j\in\Re(t_i)exp[\boldsymbol{\beta}\mathbf{X}_j]}\}^{\delta_i} \\ &= \prod_{i=1}^{D} \{\frac{exp[\sum_{k=1}^{p}\beta_k X_{(i)k}]}{\sum_{j\in\Re(t_i)exp[\sum_{k=1}^{p}\beta_k X_{jk}]}\}^{\delta^i} \end{aligned} \tag{3.80}$$

The partial likelihood is different from the usual likelihood due to the reason that it is formed by multiplying the conditional probabilities rather than multiplying the independent terms. However, it is treated as usual likelihood (Klein and Moeschberger, 2003). The numerator of the partial likelihood considers information from the individual who experiences an episode, while the denominator depends on the information from all individuals who have not yet experienced the episode or who have censored.

The partial maximum likelihood regression coefficients estimates are obtained by maximiisng the likelihood function, or, equivalently by maximising the natural logarithm of the partial likelihood function. The natural logarithm of the partial likelihood function, known as the log-partial likelihood is given by

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{D}\sum_{k=1}^{p}\beta_k X_{(i)k} - \sum_{i=1}^{D} ln[\sum_{j\in\Re(t_i)} exp(\sum_{k=1}^{p}\beta_k X_{jk})] \tag{3.81}$$

where $LL(\boldsymbol{\beta})$ denote the natural logarithm of the partial likelihood function, that is, $LL(\boldsymbol{\beta}) = ln[L(\boldsymbol{\beta})]$.

The partial derivatives of the log-partial likelihood with respect to the regression co-

efficients $(\boldsymbol{\beta})$ yields the efficient score equations (Klein and Moeschberger, 2003). The efficient score equations are given by

$$U_h(\beta) = \sum_{i=1}^{D} X_{(i)h} - \sum_{i=1}^{D} \frac{\sum_{j\in\Re(t_i)} X_{jk} exp[\sum_{k=1}^{p} \beta_k X_{jk}]}{\sum_{j\in\Re(t_i)} exp[\sum_{k=1}^{p} \beta_k X_{jk}]} \tag{3.82}$$

where $U_h(\beta)$ represent the partial derivative of the log-partial likelihood with respect to the $\beta' s$, that is, $U_h(\beta) = \frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_h}$.

The Variance estimator of the partial maximum likelihood estimates of $\boldsymbol{\beta}$ is based on the information matrix (Liu, 2012). The information matrix is obtained by taking the negative of the resulted matrix of the second derivatives of the log-partial likelihood and it is mathematically expressed as

$$\begin{aligned} I_{gh}(\boldsymbol{\beta}) = & \sum_{i=1}^{D} \frac{\sum_{j\in\Re(t_i)} X_{jg} X_{jh} exp[\sum_{k=1}^{p} \beta_k X_{jk}]}{\sum_{j\in\Re(t_i)} exp[\sum_{k=1}^{p} \beta_k X_{jk}]} \\ & - \sum_{i=1}^{D} [\frac{\sum_{j\in\Re(t_i)} X_{jg} exp(\sum_{k=1}^{p} \beta_k X_{jk})}{\sum_{j\in\Re(t_i)} exp(\sum_{k=1}^{p} \beta_k X_{jk})}] \\ & [\frac{\sum_{j\in\Re(t_i)} X_{jh} exp(\sum_{k=1}^{p} \beta_k X_{jk})}{\sum_{j\in\Re(t_i)} exp(\sum_{k=1}^{p} \beta_k X_{jk})}] \end{aligned} \tag{3.83}$$

where $I_{gh}(\boldsymbol{\beta})$ denote the second partial derivatives of the log-partial likelihood with respect to the to the $\boldsymbol{\beta' s}$, that is, $I_{gh}(\boldsymbol{\beta}) = \frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_g \beta_h}$.

Three main tests for the hypothesis about the regression coefficients $\boldsymbol{\beta}$ can be derived from these quantities. Suppose that $\mathbf{b} = (b_1, b_2, ..., b_p)'$ represent the partial maximum likelihood estimates of $\boldsymbol{\beta}$ and $\mathbf{I}(\boldsymbol{\beta}) = [I_{gh}(\beta)]_{p\times p}$ denote the information matrix assessed at $\boldsymbol{\beta}$. Under the assumption of a large sample size, **b** follows a p-variate normal distribution with mean vector $\boldsymbol{\beta}$ and the variance-covariance matrix $I^{-1}(\mathbf{b})$. The first test for the global hypothesis of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ is the wald's test and is mathematically expressed as

$$\mathbf{X}^2_{WD} = (\mathbf{b} - \boldsymbol{\beta_0})' I(\mathbf{b})(\mathbf{b} - \boldsymbol{\beta_0}) \tag{3.84}$$

which is asymptotically distributed as a chi-square with p degrees of freedom under the null hypothesis (H_0) for large sample size.

The second test for the global hypothesis of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ is the likelihood ratio test and is given by

$$\mathbf{X}^2_{LR} = 2[LL(\mathbf{b}) - LL(\boldsymbol{\beta_0})] \tag{3.85}$$

The likelihood ratio test follows a chi-square distribution with p degrees of freedom if the null hypothesis is true for large sample size. $LL(\mathbf{b})$ and $LL(\boldsymbol{\beta_0})$ denote the log-partial likelihood function containing all covariates and the log-partial likelihhod function without covariates, respectively.

The last test for the global hypothesis of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ which is based on the efficient scores is the score test and mathematically is given by

$$\begin{aligned} \mathbf{X}^2_{SC} &= \mathbf{U}(\boldsymbol{\beta_0})'\mathbf{I}^{-1}(\boldsymbol{\beta_0})\mathbf{U}(\boldsymbol{\beta_0}) \\ &= (U_1(\boldsymbol{\beta}_0), ..., U_p(\boldsymbol{\beta}_0))'\mathbf{I}^{-1}(\boldsymbol{\beta_0})(U_1(\boldsymbol{\beta}_0), ..., U_p(\boldsymbol{\beta}_0)) \end{aligned} \tag{3.86}$$

which is asymptotically distributed as chi-square with p degrees of freedom under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$. The efficient scores, $\mathbf{U}(\boldsymbol{\beta})$, is asymptotically p-variate normal with the mean of zero and variance-covariance matrix $\mathbf{I}(\boldsymbol{\beta_0})$, if H_0 is true for large samples.

3.12.1 Adjusted partial likelihood

In situations where individuals experiences episodes at the same time, that is, when there are ties between the episode times, the partial likelihood should be adjusted (Liu, 2012). Therefore, suppose that $t_1 < t_2 < .... < t_D$ denote the D ordered, distinct, episode time. If d_i is the number of episodes at time t_i and $\mathbb{D}_i$ denote the set of all patients who experiences the episode at time t_i. Let $\mathbf{Y}_i$ represent the sum of all covariates vectors $\mathbf{X}_j$ over all patients who experiences the episode at time t_i, that is, $\mathbf{Y}_i = \sum_{j \in \mathbb{D}} \mathbf{X}_j$. If $\Re(t_i)$ denote the set of all patients at risk of the event just prior time t_i, then several partial likelihood are constructed for

the data with ties among the episode times. The first partial likelihood suggested by Breslow in 1974 is called the Brewslow partial likelihood for handling ties. This likelihood works quite very well under the condition of few ties. The Breslow partial likelihood is mathematically expressed as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{exp(\boldsymbol{\beta}'\mathbf{Y}_i)}{[\sum_{j\in\Re(t_i)} exp(\boldsymbol{\beta}'\mathbf{X}_j)]^{d_i}} \tag{3.87}$$

In the Breslow partial likelihood, each of the d_i episodes at a given time are considered as distinct, creates their contribution to the likelihood function and attain the contribution of the likelihood function by taking the product over all episodes at time t_i (Klein and Moeschberger, 2003).

The second partial likelihhod due to Efron (1977) is mathematically expressed as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{exp(\boldsymbol{\beta}'\mathbf{Y}_i)}{\prod_{j=1}^{d_i}[\sum_{j\in\Re(t_i)} exp(\boldsymbol{\beta}'\mathbf{X}_j) - \frac{j-1}{d_i}\sum_{k\in\mathbb{D}_i} exp(\boldsymbol{\beta}'\mathbf{X}_k)]} \tag{3.88}$$

The Efron partial likelihood is more close to the exact partial likelihood based on the discrete hazard model as compared to the Breslow's partial likelihood. However, the two likelihood works quite the same when the number of the ties are small (Klein and Moeschberger, 2003).

3.13 Sandwich variance estimator

In recurrent survival analysis, data lines from the same individuals are treated as independent observation and therefore the regression coefficients $(\boldsymbol{\beta})$ are also estimated based on this assumption. That is, to estimate $\boldsymbol{\beta}$, the partial likelihood function $L(\boldsymbol{\beta})$ is constructed under the assumption that all observations are coming from different individuals. Using this partial likelihood function, the variance-covariance matrix of the likelihood regression coefficients estimate of $\boldsymbol{\beta}$ can be derived. However, there is a need to adjust this variance-covariance matrix as there

might be a correlation between episodes within the same patients in the study.

To estimate the variance-covariance matrix of **b**, the estimator of β, the robust sandwich variance estimator is utilised. The sandwich estimator is the widely utilised technique for adjusting the variance-covariance for the possible association among event times of the same individuals (Kleinbaum and Klein, 2006). It is important to note that the regression coefficients are not adjusted but their variances are.

The robust sandwich estimator involves the score residuals and information matrix obtained from the partial likelihood. This estimator is conveniently expressed using the matrix approach as

$$\mathbf{RS}(\hat{\beta}) = \widehat{\mathbf{Var}}(\hat{\beta})[\hat{\mathbf{R}}_s{}'\hat{\mathbf{R}}_s]\widehat{\mathbf{Var}}(\hat{\beta}) \tag{3.89}$$

where $\widehat{\mathbf{Var}}(\hat{\beta})$ is the information matrix and $\hat{\mathbf{R}}_s$ is the matrix of the score residuals.

The tests and confidence interval of global hypothesis about the regression coefficients can be conducted using this sandwich variance estimator (Kleinbaum and Klein, 2006). This can be done when considering that the estimator **b** of β follows a large sample p-variate normal distribution with the mean and variance of β and $\widehat{\mathbf{Var}}(\hat{\beta})$, respectively (Klein and Moeschberger, 2003).

3.14 Summary

This chapter presented the existing techniques of analyzing both, time to event and time to multiple events dataset. However, the time to multiple events (episodes) data analysis techniques helped to investigate the risk factors associated with recurrent episodes of peritonitis. The next chapter answers the aim and objective of the study through the results that were carried out using the recurrent survival analysis techniques presented in this chapter.

Chapter 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results of data analysis and discussion of these results. The chapter is divided into seven sections, that is, following this introductory Section 4.1, Section 4.2 looks at the descriptive statistics, Sections 4.3 and 4.4 focuses on the univariate and multivariate analysis respectively. Model selection, final multivariate model and discussion of results are presented on the following sections, 4.5, 4.6 and 4.7 respectively.

4.2 Descriptive statistics

This section gives a summary of the data set used in this study. Table 4.1 present the summary of continuous variables where the number of patients (N), mean, standard deviation (STD) and median are provided. Table 4.2 present the cate-

Table 4.1: Demographic table overall: Continuous variable

Covariate	N	Mean	STD	Median
Age	152	37.60	11.49	39
Distance (km)	138	120.33	79.25	118
Trans-Sat (%)	139	34.96	18.55	32
Ferritin (g/L)	145	491.78	335.40	448
Albumin (g/L)	145	31.83	31.15	29
GFR-MDRD (mL/min/1.73m^2)	148	6.87	3.69	6
Creatinine (mmol/L)	148	1058.17	411.30	1034.5
DBP (mmhg)	147	85.84	13.50	86
SBP (mmhg)	147	140.45	20.14	139
BMI (kg/m^2)	141	24.33	5.20	24
Hb (g/L)	148	13.51	19.11	11.1
Ca (mmol/L)	148	2.37	0.17	2.38
Pi (mmol/L)	148	1.80	2.18	1.56
K (mmol/L)	148	4.15	0.80	4
Number of people	129	4.50	1.90	4
Number of rooms	134	4.75	2.11	5

Note: The meaning of the abbreviated covariates are as follows: Trans-Sat denotes the transferrin saturation; GFR-MDRD is the glomerular filtration rate; DBP and SBP are the diastolic and systolic blood pressure, respectively; BMI depicts the body mass index; Hb is the hemoglobin content of the patient; Ca denotes the calcium, Pi is the phosphorus serum; K abbreviates potassium.

gories, frequency and percentage of the categorical variables and Table 4.3 presents the number of episodes (events) and censored values of patients. The analysis was performed based on the followed 152 PD patients, however due to censoring and other reasons, some patients did not provide all the information and therefore, the number of patients were not constant.

Table 4.1 presents the baseline clinical characteristics and lab biochemistry data of 152 PD patients. The mean and median age of patients were 38 and 39 years, respectively. The average travelled distance by patients from their home to PD centre was 120.33km, with standard deviation 0f 79.25km and median of 118km. The variables albumin and hemoglobin, had an average of 31.83g/L and 13.51g/L, standard deviation of 31.15g/L and 19.11g/L and median of 29g/L and 11.1g/L,

respectively. Mean, standard deviation and median of number of people and rooms in the homes of PD patients were 4.5, 1.9 and 4 and 4.75, 2.11 and 5, respectively.

Table 4.2: Demographic table overall: Categorical variable

Covariate	Category	Frequency	Percentage
Employment	Scholar	8	5.71
	Unemployed	109	77.86
	Employed	23	16.43
Marital status	Divorce	2	1.43
	Single	79	56.43
	Widower	1	0.71
	Married	58	41.43
Race	Black	141	92.76
	India	4	2.63
	White	7	4.61
Water	No	75	54.35
	Yes	63	45.65
Electricity	No	7	5.07
	Yes	131	94.93
Smoking	No	137	95.14
	Yes	7	4.86
Alcohol	No	142	98.61
	Yes	2	1.39
Sex	Female	73	48.03
	Male	79	51.97
Education	Primary	23	16.67
	Secondary	101	73.19
	Tertiary	14	10.14

The study population consisted of 152 PD patients who were followed in 2008 until 2012. From the 152 patients, 79 (51.97%) were males, 141 (92.76%) were blacks and 79 (56.43%) were not married (single). Table 4.2 demonstrate that many of the patients were non-smokers 139 (95.14%) , had no access to tap water 75 (54.35%), were unemployed 109 (77.86%), 131 (94.93%) had access to electricity, 142 (98.61%) did not drink alcohol, while 101 (73.19%) of these patients attended secondary school.

The study utilised the dataset which was collected at the Pietersburg provincial hospital in Limpopo province, South Africa. Limpopo is one of the province dom-

inated by black people. Hence, almost (92.76%) all of the followed patients are blacks.

Table 4.3: Summary of the number of event and censored values

Stratum	Visit	Total	Event	Censored	Event percentage	Censored percentage
1	1	152	98	54	64.47	35.53
2	2	152	55	97	36.18	63.82
3	3	152	27	125	17.76	82.24
4	4	152	9	143	5.92	94.08
5	5	152	4	148	2.63	97.37
6	6	152	1	151	0.66	99.34
Total		912	194	718	21.27	78.73

Table 4.3 shows the summary of the number of events and censored values happened during the six visits in the study. The study consisted of six strata and each stratum contained 152 patients, number of events experienced and the number of patients censored. From 912 visits there were 194 (21.27%) peritonitis episodes and 718 (78.73%) right censoring were recorded. Most of the experienced events occurred in the first stratum 98 (64.47%), followed by the second stratum with 55 (28.35%) episodes, while the last stratum constituted the least number of events 1 (0.66%). This indicates that out of 152 patients only one patient experienced the maximum of six targeted episodes of peritonitis.

The column of event percentage in Table 4.3 depicts that the number of episodes of peritonitis experienced by the PD patients decrease as the number of visits increase. To be more precise, there is an inverse relationship between the number of events (episodes) and the number of times the patient has been visited. This can be attributed to be fact that PD patients are usually trained to perform the exchange process on their own. This can be interpreted as follows, many patients pay necessary attention after experiencing the first episode.

Table 4.4: Univariate marginal model with both model-based and sandwich variance estimate for continuous clinical and social variables

			Model-Based Variance Estimate		Sandwich Variance Estimate	
Covariate	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Age	-0.0136	0.987	0.0067	0.0417	0.0099	0.1708
Distance (km)	0.0008	1.001	0.0010	0.4156	0.0012	0.5334
Trans-Sat (%)	0.0028	1.003	0.0034	0.4015	0.0067	0.6713
Ferritin (g/L)	0.0002	1.000	0.0002	0.3563	0.0003	0.5416
Albumin (g/L)	-0.0041	0.996	0.0026	0.1100	0.0044	0.3503
GFR-MDRD (mL/min/1.73m^2)	-0.0720	0.931	0.0244	0.0032	0.0307	0.0189
Creatinine (mmol/L)	0.0006	1.001	0.0002	0.0035	0.0003	0.0571
DBP (mmhg)	0.0091	1.009	0.0064	0.1538	0.0094	0.3338
SBP (mmhg)	0.0071	1.007	0.0043	0.0934	0.0066	0.2758
BMI (kg/m^2)	-0.0117	0.988	0.0142	0.4098	0.0244	0.6319
Hb (g/L)	-0.0029	0.997	0.0035	0.4108	0.0011	0.0077
Ca (mmol/L)	0.7311	2.077	0.4334	0.0916	0.5153	0.1560
Pi (mmol/L)	-0.0681	0.934	0.0921	0.4599	0.0775	0.3800
K (mmol/L)	-0.2014	0.818	0.1085	0.0634	0.1409	0.1528
Number of people	0.0617	1.064	0.0385	0.1095	0.0499	0.2166
Number of rooms	-0.0416	0.959	0.0403	0.3022	0.0530	0.4325

Note: The meaning of the abbreviated covariates are as follows: Trans-Sat denotes the transferrin saturation; GFR-MDRD is the glomerular filtration rate; DBP and SBP are the diastolic and systolic blood pressure, respectively; BMI depicts the body mass index; Hb is the hemoglobin content of the patient; Ca denotes the calcium, Pi is the phosphorus serum; K abbreviates potassium.

4.3 Univariate analysis

The results of the univariate models with both sandwich and model-based variance estimate, for the continuous covariates are presented from Table 4.4 until Table 4.11 The term univariate indicates that covariates in the data set were first fitted one by one separately in both the sandwich and model-based variance models. The parameter estimates from both sandwich and model-based variance models are the same. The only difference is in the standard errors of the parameter estimates. The sandwich variance model takes into account the correlation structure among the recurrent events per person but the model based variance estimate does not. Certainly, due to the change in the standard error of the parameter estimates coming out of the two modelling approaches, their p-values also will change accordingly. The lower standard error will result in a lower p-value and vice versa.

Patient age at baseline, GFR-MDRD and Creatinine are significant covariates in the model-based univariate marginal survival model with the p-values, 0.0147, 0.0032 and 0.0035, respectively. However, when the correlation for the recurrent events among the same patient is taken into account by the sandwich variance marginal model, only covariates GFR-MDRD and haemoglobin content (Hb) are significant with the p-values 0.0189 and 0.0077, respectively. The regression co-

efficient of GFR-MDRD is -0.0720, with the hazard ratio of 0.931, indicating that for each unit increase on GFR-MDRD content the hazard of experiencing recurrent episodes of peritonitis decreases by 6.9%. A less than one hazard ratio is obtained for the covariate haemoglobin content (HR=0.997). Each unit increase in haemoglobin content resulted in reducing the risk of experiencing recurrent episode of peritonitis by 0.3%. However, the result obtained from both of the univariate modelling approaches revealed that, covariates distance from the dialysis centre, transferrinsat, ferritin, albumin, DBP, SBP, BMI, Ca, Pi, K, number of people and number of rooms are statistically non-significant at 5% level of significance (Table 4.4).

Table 4.5: Univariate counting process model with both model-based and sandwich variance estimate for continuous clinical and social variables

			Model-Based Variance Estimate		Sandwich Variance Estimate	
Covariate	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Age	-0.0104	0.990	0.0067	0.1207	0.0072	0.1505
Distance (km)	0.0006	1.001	0.0010	0.5172	0.0010	0.5109
Trans-Sat (%)	0.0034	1.003	0.0034	0.3690	0.0059	0.5638
Ferritin (g/L)	0.0002	1.000	0.0002	0.3876	0.0003	0.4480
Albumin (g/L)	-0.0033	0.997	0.0027	0.2277	0.0042	0.4374
GFR-MDRD (mL/min/1.73m^2)	-0.0534	0.948	0.0242	0.0273	0.0219	0.0148
Creatinine (mmol/L)	0.0006	1.000	0.0002	0.0392	0.0002	0.0756
DBP (mmhg)	0.0047	1.005	0.0064	0.4589	0.0074	0.5244
SBP (mmhg)	0.0060	1.006	0.0047	0.1701	0.0053	0.2579
BMI (kg/m^2)	-0.0043	0.996	0.0140	0.7578	0.0166	0.7951
Hb (g/L)	-0.0020	0.998	0.0035	0.5715	0.0009	0.0313
Ca (mmol/L)	0.6846	1.983	0.4652	0.1411	0.4138	0.0980
Pi (mmol/L)	-0.0565	0.945	0.0888	0.5248	0.0560	0.3123
K (mmol/L)	-0.1395	0.870	0.1125	0.2151	0.1171	0.2338
Number of people	0.0449	1.046	0.0389	0.2483	0.0383	0.2409
Number of rooms	-0.0323	0.968	0.0395	0.4130	0.0393	0.4107

Note: The meaning of the abbreviated covariates are as follows: Trans-Sat denotes the transferrin saturation; GFR-MDRD is the glomerular filtration rate; DBP and SBP are the diastolic and systolic blood pressure, respectively; BMI depicts the body mass index; Hb is the hemoglobin content of the patient; Ca denotes the calcium, Pi is the phosphorus serum; K abbreviates potassium.

The regression coefficients of GFR-MDRD and Creatinine are statistically significant at 5% level of significance in the model-based univariate counting process model with p-values 0.0273 and 0.0392, respectively. However, when the correlation for the recurrent events among same patient is adjusted through the sandwich variance estimator, only regression coefficient of GFR-MDRD and hemoglobin content are statistically significant with p-values 0.0148 and 0.0313, respectively. The regression coefficient of GFR-MDRD is -0.0534 and the hazard ratio obtained by exponentiating this regression coefficient is 0.948. This hazard ratio suggests that each unit increase in GFR-MDRD lowers the rate of experiencing recurrent

episodes of peritonitis by about 5%. The hazard ratio obtained by exponentiating the regression coefficient (-0.0020) of hemoglobin content is 0.998 and this value indicates that each unit increase in hemoglobin content result in lowering the rate of experiencing recurrent episodes of peritonitis by 0.2%. Distance from dialysis center, transferritinsat, ferritin, albumin, DBP, SBP, BMI, Ca, Pi, K, number of people and rooms are statistically non-significant at 5% level of significance in both model-based and sandwich variance models.

Table 4.6: Univariate gap-time model with both model-based and sandwich variance estimate for continuous clinical and social variables

			Model-Based Variance Estimate		Sandwich Variance Estimate	
Covariate	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Age	-0.0110	0.989	0.0067	0.1019	0.0068	0.1055
Distance (km)	0.0008	1.001	0.0010	0.3945	0.0009	0.3445
Trans-Sat (%)	0.0034	1.004	0.0040	0.3657	0.0057	0.5309
Ferritin (g/L)	0.0000	1.000	0.0002	0.8204	0.0002	0.8339
Albumin (g/L)	-0.0023	0.998	0.0028	0.4092	0.0043	0.5900
GFR-MDRD (mL/min/1.73m^2)	-0.0413	0.960	0.0239	0.0836	0.0218	0.0573
Creatinine (mmol/L)	0.0003	1.000	0.0002	0.1193	0.0002	0.1847
DBP (mmhg)	0.0039	1.004	0.0066	0.5551	0.0078	0.6208
SBP (mmhg)	0.0055	1.006	0.0045	0.2139	0.0054	0.3053
BMI (kg/m^2)	-0.0066	0.993	0.0143	0.6426	0.0144	0.6461
Hb (g/L)	-0.0016	0.998	0.0035	0.6414	0.0010	0.0970
Ca (mmol/L)	0.4616	1.587	0.4667	0.3226	0.4498	0.3048
Pi (mmol/L)	-0.0612	0.941	0.0923	0.5076	0.0565	0.2786
K (mmol/L)	-0.1346	0.874	0.1135	0.2357	0.1131	0.2340
Number of people	0.0280	1.028	0.0402	0.4859	0.0372	0.4513
Number of rooms	-0.0364	0.964	0.0395	0.3570	0.0361	0.3136

Note: The meaning of the abbreviated covariates are as follows: Trans-Sat denotes the transferrin saturation; GFR-MDRD is the glomerular filtration rate; DBP and SBP are the diastolic and systolic blood pressure, respectively; BMI depicts the body mass index; Hb is the hemoglobin content of the patient; Ca denotes the calcium, Pi is the phosphorus serum; K abbreviates potassium.

Results of the univariate gap-time model both with sandwich and model-based variance estimate for continuous covariates presented in Table 4.6 revealed that the regression coefficients of all the fitted covariates are statistically non-significant at 5% level of significance. That is, the p-values corresponding to all regression coefficients are greater than 0.05 level of significance.

Covariates GFR-MDRD and creatinine are statistically significant at 5% level of significance in the model-based univariate stratified model with p-values 0.0340 and 0.0435, respectively. The covariate creatinine is marginally significant and this results into non-significant when the correlation for recurrent events among the same patient is taken into account by the sandwich variance stratified model. The standard error of GFR-MDRD is small in the sandwich than in the model-based stratified model and this results in small p-value (0.0142), making the covariate

Table 4.7: Univariate stratified model with both model-based and sandwich variance estimate for continuous clinical and social variables

			Model-Based Variance Estimate		Sandwich Variance Estimate	
Covariate	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Age	-0.0100	0.990	0.0069	0.1447	0.0064	0.1161
Distance (km)	0.0005	1.001	0.0010	0.5907	0.0009	0.5465
Trans-Sat (%)	0.0044	1.004	0.0039	0.2572	0.0058	0.4421
Ferritin (g/L)	0.0002	1.000	0.0002	0.3971	0.0002	0.3847
Albumin (g/L)	-0.0027	0.997	0.0029	0.3568	0.0042	0.5205
GFR-MDRD (mL/min/1.73m^2)	-0.0529	0.948	0.0250	0.0340	0.0216	0.0142
Creatinine (mmol/dL)	0.0004	1.000	0.0002	0.0435	0.0002	0.0582
DBP (mmhg)	0.0048	1.005	0.0066	0.4622	0.0072	0.5011
SBP (mmhg)	0.0055	1.005	0.0044	0.2194	0.0050	0.2784
BMI (kg/m^2)	-0.0098	0.990	0.0150	0.5130	0.0154	0.5251
Hb (g/L)	-0.0020	0.998	0.0035	0.5707	0.0011	0.0603
Ca (mmol/L)	0.5732	1.774	0.4772	0.2297	0.3948	0.1466
Pi (mmol/L)	-0.0364	0.964	0.0773	0.6371	0.0449	0.4165
K (mmol/L)	-0.0762	0.927	0.1135	0.5021	0.1105	0.4906
Number of people	0.0410	1.042	0.0414	0.3215	0.0369	0.2666
Number of rooms	-0.0299	0.971	0.4582	0.4582	0.0352	0.3964

Note: The meaning of the abbreviated covariates are as follows: Trans-Sat denotes the transferrin saturation; GFR-MDRD is the glomerular filtration rate; DBP and SBP are the diastolic and systolic blood pressure, respectively; BMI depicts the body mass index; Hb is the hemoglobin content of the patient; Ca denotes the calcium, Pi is the phosphorus serum; K abbreviates potassium.

to remain statistically significant at 5% level of significance in the sandwich variance stratified model. The regression coefficient of GFR-MDRD is -0.0529, with the hazard ratio of 0.931, suggesting that each unit increase in GFR-MDRD results in reducing the rate of experiencing recurrent episodes of peritonitis by about 7%. The regression coefficients of the other fitted covariates are statistically non-significant at 5% level of significance in both model-based and sandwich variance stratified models.

Summary

Based on three of the four fitted univariate regression models: marginal, stratified and counting process, the glomerular filtration rate was consistently significant at 5% level of significance. Hemoglobin content was found to be significant only in the counting process model. No continuous covariate was found to be significant when using the univariate gap-time regression model. Moreover, hemoglobin content and glomerular filtration rate were the only two covariates qualifying to be taken to the multivariate regression model. This can be supported by fact that, they were the only significant covariates in one or more fitted univariate models.

In Table 4.8, the regression coefficients of unemployed patients and patients with-

Table 4.8: Univariate marginal model with both model-based and sandwich variance estimate for categorical clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Employment	Scholar	-0.7503	0.472	0.6143	0.2220	0.5899	0.2034
	Unemployed	0.4865	1.627	0.2202	0.0272	0.2744	0.0763
	Employed(ref)	0.0000	1.000				
Marital status	Divorce	-14.0213	0.000	468.0413	0.9761	0.8663	$<$0.0001
	Single	0.1153	1.122	0.1534	0.4522	0.2398	0.6305
	Widower	0.5195	1.681	0.7207	0.4710	0.2240	0.0204
	Married(ref)	0.0000	1.000				
Race	Black	0.8846	2.422	0.5065	0.0807	0.7246	0.2222
	India	-0.1076	0.898	0.7087	0.8793	1.1414	0.9249
	White(ref)	0.0000	1.000				
Water	No	0.3670	1.443	0.1610	0.0226	0.2304	0.1113
	Yes(ref)	0.0000	1.000				
Electricity	No	0.3352	1.398	0.3129	0.2841	0.3048	0.2714
	Yes(ref)	0.0000	1.000				
Smoking	No	0.0096	1.010	0.3265	0.9766	0.3129	0.9756
	Yes(ref)	0.0000	1.000				
Sex	Female	0.1473	1.159	0.1452	0.3102	0.2169	0.4970
	Male(ref)	0.0000	1.00				
Education	Primary	0.4247	1.529	0.3168	0.1801	0.4622	0.3582
	Secondary	0.0865	1.090	0.2731	0.7514	0.3547	0.8073
	Tertiary(ref)	0.0000	1.000				

out access to tap water are statistically significant at 5% level of significance in the model-based marginal model. However, when the correlation among recurrent events of the same patients is adjusted by the robust sandwich variance estimator, the covariate marital status is statistically significant at 5% level of significance. Divorce and widower are the significant categories of marital status with p-values $<$0.0001 and 0.0204, respectively. The hazard ratio of 0.000 for divorce suggests that the risk of recurrent episodes of peritonitis in divorced patients is about 100% lower than among those who are married. In Table 4.2, it is reported that the number of divorced patients were only 2 out of the 152 followed patients. Hence, this hazard ratio is not of great concern. The regression coefficient of widower is 0.5195, with hazard ratio 1.681, indicating that about 68% rate of experiencing recurrent episodes of peritonitis for widower as compared to married patients. Race of patients, sex, education status, smoking status and access to electricity are consistently non-significant covariates at 5% level of significance in both model-based and sandwich variance models.

In the univariate counting process model with model-based variance estimate, all the fitted categorical covariates are statistically non-significant at 5% level of sig-

Table 4.9: Univariate counting process model with both model-based and sandwich variance estimate for categorical clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Employment	Scholar	-0.5477	0.578	0.6131	0.3717	0.5069	0.2799
	Unemployed	0.3358	1.399	0.2201	0.1271	0.2210	0.1286
	Employed(ref)	0.0000	1.000				
Marital status	Divorce	-14.0190	0.000	677.8823	0.9835	0.7889	$<$0.0001
	Single	0.1019	1.107	0.1593	0.5222	0.1496	0.4956
	Widower	0.4472	1.564	0.7321	0.5413	0.2003	0.0256
	Married(ref)	0.0000	1.000				
Race	Black	0.5350	1.707	0.5101	0.2943	0.6566	0.4152
	India	0.0162	1.016	0.7103	0.9818	0.9251	0.9860
	White(ref)	0.0000	1.000				
Water	No	0.1600	1.173	0.1668	0.3376	0.1708	0.3489
	Yes(ref)	0.0000	1.000				
Electricity	No	0.1896	1.209	0.3174	0.5502	0.1927	0.3250
	Yes(ref)	0.0000	1.000				
Smoking	No	0.0026	1.003	0.3310	0.9937	0.1966	0.9894
	Yes(ref)	0.0000	1.000				
Sex	Female	0.0106	1.011	0.1500	0.9436	0.1445	0.9415
	Male(ref)	0.0000	1.00				
Education	Primary	0.2784	1.321	0.3272	0.3950	0.3339	0.4045
	Secondary	-0.0347	0.966	0.2777	0.9007	0.2883	0.9043
	Tertiary(ref)	0.0000	1.000				

nificance. Nevertheless, when the correlation among recurrent events of the same patients is accounted by the sandwich variance estimator, covariate marital status is statistically significant at 5% level of significance. The regression coefficients of categories divorce and widower are -14.0190 and 0.4472, respectively. The hazard ratio (HR=1.564) of more than one is obtained by exponentiating the regression coefficient of widower. The meaning of this coefficient is that the rate of widower to experience recurrent episodes of peritonitis is two times higher among those who are married. The other fitted covariates are also statistically non-significant in 5% level of significance even in the robust sandwich univariate counting process model.

Results of univariate gap-time model with both model-based and sandwich variance estimate for categorical covariates presented in Table 4.10 reveals that all the fitted covariates are statistically non-significant at 5% level of significance in the model-based variance model. The regression coefficient of category divorce of covariate marital status is negatively and statistically significant at 5% level of significance ($\beta = -13.0362$,p-value$<$0.0001) when the correlation among recurrent episodes of the same patient is adjusted by the sandwich variance estimate.

Table 4.10: Univariate gap-time model with both model-based and sandwich variance estimate for categorical clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Employment	Scholar	-0.5077	0.602	0.6181	0.4114	0.5484	0.3546
	Unemployed	0.3123	1.367	0.2233	0.1620	0.2251	0.1653
	Employed(ref)	0.0000	1.000				
Marital status	Divorce	-13.0362	0.000	408.7351	0.9746	0.7806	<0.0001
	Single	0.0712	1.0742	0.1553	0.6467	0.1535	0.6428
	Widower	0.2536	1.289	0.7230	0.7257	0.4211	0.5470
	Married(ref)	0.0000	1.000				
Race	Black	0.5431	1.721	0.5097	0.2866	0.5457	0.3196
	India	-0.2657	0.767	0.7156	0.7104	0.6996	0.7041
	White(ref)	0.0000	1.000				
Water	No	0.1694	1.185	0.1634	0.2997	0.1627	0.2978
	Yes(ref)	0.0000	1.000				
Electricity	No	0.1694	1.206	0.3144	0.5522	0.2732	0.4938
	Yes(ref)	0.0000	1.000				
Smoking	No	0.1537	1.166	0.3325	0.6438	0.2331	0.5096
	Yes(ref)	0.0000	1.000				
Sex	Female	0.0217	1.022	0.1477	0.8831	0.1448	0.8808
	Male(ref)	0.0000	1.00				
Education	Primary	0.2457	1.279	0.3187	0.4407	0.3344	0.4625
	Secondary	-0.0676	0.935	0.2763	0.8068	0.3001	0.8218
	Tertiary(ref)	0.0000	1.000				

The risk of experiencing recurrent episodes of peritonitis for divorced patients is 100% lower than among those who are married. Covariates Employment status, race, smoking status, sex, education, water and electricity are consistently non-significant at 5% level of significance in the both approaches.

Table 4.11: Univariate stratified model with both model-based and sandwich variance estimate for categorical clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Employment	Scholar	-0.4452	0.642	0.6194	0.4722	0.5486	0.4171
	Unemployed	0.3474	1.415	0.2251	0.1227	0.2152	0.1064
	Employed(ref)	0.0000	1.000				
Marital status	Divorce	-14.0190	0.000	677.8823	0.9835	0.7891	<0.0001
	Single	0.1019	1.107	0.1593	0.5222	0.1557	0.5132
	Widower	0.4472	1.564	0.7321	0.5413	0.3954	0.2581
	Married(ref)	0.0000	1.000				
Race	Black	0.5350	1.707	0.5101	0.2943	0.5408	0.3226
	India	0.0162	1.016	0.7282	0.9822	1.1414	0.9249
	White(ref)	0.0000	1.000				
Water	No	0.1560	1.173	0.1668	0.3349	0.1659	0.3349
	Yes(ref)	0.0000	1.000				
Electricity	No	0.1896	1.209	0.3174	0.5502	0.2800	0.4983
	Yes(ref)	0.0000	1.000				
Smoking	No	0.0026	1.001	0.3310	0.9937	0.2544	0.9918
	Yes(ref)	0.0000	1.000				
Sex	Female	0.0106	1.011	0.1500	0.9436	0.1462	0.9421
	Male(ref)	0.0000	1.00				
Education	Primary	0.2784	1.321	0.3272	0.3950	0.3439	0.4183
	Secondary	-0.0347	0.966	0.2777	0.9007	0.3004	0.9081
	Tertiary(ref)	0.0000	1.000				

Table 4.11 reveals that all the fitted covariates are statistically non-significant at 5%

level of significant in the stratified univariate model-based variance estimate. Divorce is the significant categories of covariate marital status with p-values $<$0.0001. The hazard ratio of 0.000 for divorce suggests that the rate of divorced patients to experience recurrent episodes of peritonitis is about 100% lower than among those who are married. The other seven fitted covariates are also statistically non-significant even in the robust sandwich univariate stratified model.

Summary

Based on the results obtained from the four univariate regression models, it can be concluded that marital status is the only covariates selected to be included in the final multivariate regression models. That is, marital status is the only significant covariate at 5% level. Moreover, the regression coefficient of the divorce category of the marital status is very high, which makes the hazard ratio to be very small. This can be attributed to the fact that only 2 (0.01%) PD patients were divorced and compared to 58 (0.38%) married PD patients.

4.4 Multivariate analysis

The multivariate models for the four recurrent survival analysis techniques (marginal, counting process, gap-time and stratified model) with both model-based and sandwich variance estimates are fitted and presented from Table 4.12 until Tablev4.15. The covariates that were statistically significant at 5% level of significant in the univariate sandwich variance models are fitted together in the same model, thus the model is referred to as the multivariate model. The regression coefficients of covariates from both sandwich and model-based variance marginal models are the same. However, these regression coefficients have different standard error and due to the change in the standard error of the regression coefficients coming out of the two modelling approaches, their p-values will also change accordingly. That

is, lower standard error will result in a lower p-value and vice versa.

Table 4.12: Multivariate marginal model with both model-based and sandwich variance estimate for both clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Marital status	Divorce	-14.2832	0.000	475.0718	0.9760	0.9151	$<$0.0001
	Single	0.0509	1.052	0.1559	0.7440	0.2499	0.8385
	Widower	0.6400	1.896	0.7229	0.3760	0.2220	0.0039
	Married(ref)	0.0000	1.000				
GFR-MDRD (mL/min/1.73m^2)		-0.1049	0.900	0.0281	0.0002	0.0369	0.0045
Hb (g/dL)		-0.0031	0.997	0.0034	0.3542	0.0010	0.0019

Note: The meaning of the abbreviated covariates are as follows: GFR-MDRD is the glomerular filtration rate; Hb is the hemoglobin content of the patient.

The parameter estimate of GFR-MDRD is the only statistically significant parameter in the model-based multivariate marginal model with p-value 0.0002. However, when the correlation for the recurrent events among the same patient is adjusted by the sandwich variance estimator, covariate marital status categories, divorce and widower, GFR-MDRD and hemoglobin are statistically significant at 5% level of significance with the p-values $<$0.0001, 0.0039, 0.0045, and 0.0019, respectively.

The regression coefficient of divorce is -14.2832, with the hazard ratio of 0.000, indicating 100% lower risk of experiencing recurrent episodes of peritonitis in divorced patients as compared to married patients. The hazard ratio estimate of 1.052 for the parameter estimate of widower indicates that the rate of experiencing recurrent episodes of peritonitis for widower is 5.2% higher than the married patients. The hazard estimate of 0.990 suggests that each unit increase in GFR-MDRD lowers the risk of experiencing recurrent episodes of peritonitis by about 10%. Negative regression coefficient (-0.0038) for Hb generates the hazard ratio less than one (HR=0.997, that is, each unit increases in hemoglobin content reduces the rate of experiencing recurrent episodes of peritonitis by about 0.3%.

Table 4.13: Multivariate counting process model with both model-based and sandwich variance estimate for both clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Marital status	Divorce	-14.1444	0.000	570.8863	0.9802	0.8566	$<$0.0001
	Single	0.0356	1.036	0.1550	0.8181	0.1698	0.8338
	Widower	0.3161	1.372	0.7179	0.6605	0.1386	0.0226
	Married(ref)	0.0000	1.000				
GFR-MDRD (mL/min/1.73m^2)		-0.0712	0.931	0.0268	0.0079	0.0254	0.0050
Hb (g/dL)		-0.0020	0.998	0.0033	0.5448	0.0008	0.0147

Note: The meaning of the abbreviated covariates are as follows: GFR-MDRD is the glomerular filtration rate; Hb is the hemoglobin content of the patient.

Results of the multivariate counting process model with model-based variance estimate reveals that GFR-MDRD with the regression coefficient of -0.0712 is the

only significant covariate at 5% level of significance. However, when the variance structure is adjusted by the robust sandwich variance estimate, marital status categories, divorce and widower, and GFR-MDRD are the only significant covariates. The regression coefficient of widower is 0.3161, with the hazard ratio of 1.372, suggesting that recurrent episodes of peritonitis rate in divorced patients is about 40% times higher than among those who are married. The regression coefficient of GFR-MDRD is -0.0712 and exponentiating this value generates the hazard ratio of 0.931 which indicates about 7% lower risk of experiencing recurrent episodes of peritonitis when GFR-MDRD is increased by one unit. The hazard ratio estimate of 0.992 for the parameter estimate of Hb indicates that for each unit increase in Hb the rate of experiencing recurrent episodes of peritonitis decreases by 0.2%.

Table 4.14: Multivariate gap-time model with both model-based and sandwich variance estimate for both clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Marital status	Divorce	-13.2059	0.000	415.9706	0.9747	0.8271	$<$0.0001
	Single	0.0354	1.036	0.1594	0.8244	0.1634	0.8336
	Widower	0.3646	1.440	0.7253	0.6152	0.1535	0.0176
	Married(ref)	0.0000	1.000				
GFR-MDRD (mL/min/1.73m^2)		-0.0614	0.940	0.0271	0.0236	0.0259	0.0178
Hb (g/L)		-0.0013	0.999	0.0034	0.6930	0.0010	0.1742

Note: The meaning of the abbreviated covariates are as follows: GFR-MDRD is the glomerular filtration rate; Hb is the hemoglobin content of the patient.

The covariate GFR-MDRD is the only significant covariate in the model-based multivariate gap-time model with p-value of 0.0236. However, when the correlation for recurrent events among the same patients is handled by the sandwich variance model, covariates marital status categories, divorce and widower, and GFR-MDRD are statistically significant at 5% level of significance with p-values $<$0.0001, 0.0176 and 0.0178, respectively.

The hazard ratio (HR=0.000) of the regression coefficient (β=-13.3385) for divorced patients indicates 100% lower rate of experiencing recurrent episodes of peritonitis than among patients who are married; the hazard estimate of marital status category, single, suggests that the rate of experiencing recurrent episodes is 44% higher as compared to married patients. The regression coefficient of GFR-MDRD is -0.0558, with the hazard ratio of 0.940, suggesting that the rate of experiencing recurrent episodes of peritonitis decreases by 6% when GFR-MDRD increases by

one unit.

Both modelling approaches revealed that patient's employment status and hemoglobin content are statistically non-significant covariates at 5% level of significance.

Table 4.15: Multivariate stratified model with both model-based and sandwich variance estimate for both clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Marital status	Divorce	-14.2096	0.000	684.7926	0.9834	0.8298	$<$0.0001
	Single	0.0694	1.072	0.1610	0.6666	0.1530	0.6503
	Widower	0.5473	1.729	0.7339	0.4558	0.1959	0.0052
	Married(ref)	0.0000	1.000				
GFR-MDRD (mL/min/1.73m^2)		-0.0720	0.931	0.0279	0.0099	0.0258	0.0053
Hb (g/L)		-0.0020	0.998	0.0034	0.5652	0.0011	0.0648

Note: The meaning of the abbreviated covariates are as follows: GFR-MDRD is the glomerular filtration rate; Hb is the hemoglobin content of the patient.

The regression coefficient of the covariate GFR-MDRD is negatively and statistically significantly associated with recurrent episodes of peritonitis in the model-based stratified model (β=-0.0720,P=0.0099). However, when the correlation for the recurrent events among the same patient is taken into account by the sandwich variance stratified model, only covariates marital status 'divorce and widower', GFR-MDRD and hemoglobin content are statistically significant at 5% level of significant with p-values $<$0.0001, 0.0052, and 0.0053, respectively.

The rate of experiencing recurrent episodes of peritonitis for divorced patients is lower by 100% (HR=0.000) than among those who are married. The hazard ratio of greater than one is obtained for the covariate marital status category widower (HR=1.729). That is, the rate of widower to experience recurrent episodes of peritonitis is about twice higher than among in the married patient. The regression coefficient of GFR-MDRD is -0.0720, with the hazard ratio of 0.932, suggesting that each unit increase in GFR-MDRD reduces the rate of experiencing recurrent episodes of peritonitis by about 7%.

Summary

The four multivariate techniques for handling recurrent events dataset discussed in chapter 3 are fitted in this section. The marginal and counting process models has demonstrated that all the fitted covariates (marital status, GFR-MDRD and Hb

content) are statistically significant at the level of 5%. Moreover, the gap-time and marginal models could not say the same thing about the Hb content. To be precise, marital status and GFR-MDRD were also significant in the gap-time and marginal regression model, while the hemoglobin (Hb) content was not.

4.5 Model selection

Four recurrent survival analysis techniques are fitted to check the significance of the risk factors associated with peritonitis. Significant factors are identified from the four techniques, however, the aim of the study is to identify factors and compare the techniques as for which one fit the dataset better than the other models. Table 4.16 below present the two model comparison techniques used to assess the fitness of the models.

Table 4.16: Model comparison

	Criterion	
Model	AIC	SBC
Marginal	1365.017	1380.892
Counting Process	1491.361	1507.241
Gap-time	1220.946	1236.770
Stratified	1136.254	1152.134

The AIC and SBC comparison methods are utilised in order to select the best model from the four fitted recurrent survival models. Therefore, the stratified proportional hazard model has the smallest value of AIC and SBC as compared to the other three methods, 1136.254 and 1152.134, respectively, as reflected in Table 4.16. Kleinbaum and Klein (2006) stated that a model with the smallest value is more preferable and this is enough to conclude that the stratified proportional hazard regression model is the best technique for fitting the available recurrent episodes data set.

4.6 Final multivariate model

This study utilized a small sample size data set to fit the models and due to this the small number of ties was observed between the episodes times. Kleinbaum and Klein (2006) indicated that when the number of ties among individual's episodes times are small the Breslow and Efron partial likelihoods work quite the same and this correspond with what was perceived in this study. The following results are obtained using the Breslow method of handling ties, however, due to the small number of ties among the episodes time these results are parallel to the ones obtained when utilizing the Efron method of handling ties.

Table 4.17: Testing global null hypothesis: $\mathbf{H}_0 : \boldsymbol{\beta} = 0$

Test	Chi-square	DF	P-value
Likelihood Ratio	14.3142	5	0.0137
Score(Model-based)	10.6242	5	0.0594
Score(Sandwich)	12.9849	5	0.0235
Wald(Model-based)	7.6191	5	0.1785
wald(Sandwich)	321.4213	5	$<$0.0001

In this model there are three extra tests for the global null hypothesis. The first test is the likelihood ratio test and it is statistically significant at 5% level of significance with the p-value of 0.0137. The score chi-square statistic based on the robust sandwich variance estimator is not much larger than the chi-square statistic from the model-based variance estimator. However, the model-based test statistic is not significant at 5% level of significance. The wald test statistic based on the robust sandwich variance estimator is much large than the model-based statistics. Furthermore, it is extremely significant at 5% level of significance, while the model-based is non-significant on the same level of significance.

The standard errors based on the robust sandwich variance estimate for Hb, GFR-MDRD and marital status categories, single and widower, are slightly larger than the model-based variance with the standard error of the parameter estimate of

Table 4.18: Multivariate stratified model with both sandwich and model-based variance estimate for both clinical and social variables

				Model-based variance estimate		Sandwich variance estimate	
Covariate	Category	Parameter Estimate	Hazard Ratio	Standard Error	P-Value	Standard Error	P-Value
Marital status	Divorce	-14.2096	0.000	684.7926	0.9834	0.8298	$<$0.0001
	Single	0.0694	1.072	0.1610	0.6666	0.1530	0.6503
	Widower	0.5473	1.729	0.7339	0.4558	0.1959	0.0052
	Married(ref)	0.0000	1.000				
GFR-MDRD (mL/min/1.73m^2)		-0.0720	0.931	0.0279	0.0099	0.0258	0.0053
Hb (g/dL)		-0.0020	0.998	0.0034	0.5652	0.0011	0.0648

Note: The meaning of the abbreviated covariates are as follows: GFR-MDRD is the glomerular filtration rate; Hb is the hemoglobin content of the patient.

marital status category, divorce, being extremely larger.

The regression coefficient for GFR-MDRD, marital status categories, divorced and widower, are statistically significant at 5% level of significance with p-values 0.0053, $<$0.0001, and 0.0052, respectively.

The hazard ratio estimate of 0.000 for divorced patients indicates 100% lower rate of experiencing recurrent episodes of peritonitis when compared to married patients; the hazard ratio for single and widower indicates 7.2 and 72.9 higher chances of experiencing recurrent episodes of peritonitis, respectively as compared to married patients. The hazard ratio estimates for GFR-MDRD and Hb are 0.931 and 0.998, respectively, suggesting that for each unit increase in GFR-MDRD and Hb the rate of experiencing recurrent episodes of peritonitis decreases by around 7% and 0.2%, respectively.

Summary

The stratified proportional hazard model is selected to be the better fitting model for the available dataset of recurrent episodes of peritonitis. The selection was made by considering the proportional regression model with the smallest AIC and SBC amongst the four competing models.

The global null hypothesis of whether there exist some risk factors associated with recurrent episodes of peritonitis was done using likelihood, score and wald test procedures. The test statistic value for the score and wald test were calculated for both model based and sandwich variance estimate. Moreover, the results obtained when the model based variance estimator was considered suggested that there is

no risk factors associated with recurrent episodes of peritonitis. The results based on the sandwich variance estimator contradicted the results from the model based variance structure. That is, all the three testing procedures depicted that there are some significant risk factors associated with recurrent episodes of peritonitis.

The significant risk factors associated with recurrent episodes of peritonitis were investigated using the selected better fitting stratified proportional hazard model. Marital status of the followed PD patients and GFR-MDRD were found to be the major significant risk factors of recurrent episodes of peritonitis at the level of 5%.

4.7 Discussion

There have been several studies for investigating the potential major risk factors influencing the occurrence of peritonitis in PD patients. However, majority of them focused on examining the risk factors associated with time-to-first episode of peritonitis. This study focuses in comparing different recurrent survival analysis techniques and use the better fitting technique to investigate the potential risk factors associated with recurrent episodes of peritonitis.

Univariate counting process, stratified, gap-time and marginal hazard regression models are performed to select significant covariates to the multivariate regression hazard models. Regression coefficient for covariates are considered statistically significant when their corresponding p-values were less than 5%. Four multivariate hazard regression models are fitted with the selected covariates and compared using the AIC and SBC value, as for which one is the better technique to fit recurrent episodes (events) data set. Model with the smallest AIC and SBC value is considered to be the good fitting model for the available recurrent episodes data set. The major risk factors associated with recurrent episodes peritonitis are examined from the selected good fitting model.

The stratified regression hazard model is the one with the smallest AIC and SBC

value when compared to the other three recurrent modelling techniques. Therefore, this is considered as enough evidence to conclude that stratified regression hazard model is the better technique for fitting the recurrent events (peritonitis episodes) data set. This findings are inconsistent with what was declared in 2003 by Duchateau and his colleagues. (Duchateau et al., 2003) stated the gap-time modelling technique is the most suitable approach for studying recurrent or multiple events data set. However, in this study, the gap-time hazard regression model appears to be the second best approach for fitting recurrent episodes data set.

The study identified two independent risk factors to be significantly associated with recurrent episodes of peritonitis: marital status and glomerular filtration rate. Two categories of marital status, that is, divorce and widower, are the significant factors when compared to married patients (when taking married patients as the reference category). These findings are consistent with the results produced by the gap-time modelling approach. Hemoglobin content is found to be one of the risk factors for peritonitis in the study conducted by Isla et al. (2014). However, in this study hemoglobin is found to be statistically significant in the counting process and marginal hazard regression models and not statistically significant in the stratified and gap-time modelling techniques. Therefore, it is considered to be not a potential risk factors because is not significant in the best fitting model.

The study also revealed that single and widower patients when compared to married patients have high risk rate of experiencing recurrent episodes of peritonitis. This findings suggest that health practitioners must pay more attention to single and widower patients when giving instructions of how to operate the peritoneal dialysis treatment. In addition, the study revealed that a unit increase in glomerular filtration rate and hemoglobin content reduces the rate of experiencing recurrent episodes of peritonitis. This findings suggest that the lower the glomerular filtration rate and hemoglobin content in PD patients, the worse the rate of experiencing recurrent episodes of peritonitis. Therefore, it is important for doctors to monitor the patients's glomerular filtration rate and hemoglobin contents, and also to give them

advices on how to keep their contents high so that they can avoid this infection.

Despite these significant discoveries, this study have some limitations which might have strongly influenced the results. These limitations highlight the significance of future research on comparison of recurrent survival analysis techniques and use them to identify the independent major risk factors associated with recurrent episodes of peritonitis.

Chapter 5

CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This is the last chapter of the study. This chapter focuses on summarizing the previous four chapters. Conclusion based on the findings allied to the aim and objectives of the study is also discussed. The study limitations, areas for further studies and recommendations are outlined in this chapter as well.

5.2 Summary and research findings

The aim of this study was to investigate the major risk factors associated with recurrent episodes of peritonitis among the kidney patients who are in peritoneal dialysis at the Pietersburg provincial hospital in Limpopo, South Africa. This aim

was investigated using the selected best fitting recurrent survival analysis technique. More specifically, four recurrent survival analysis techniques were fitted on the multiple (recurrent) episodes of peritonitis dataset. The four techniques were compared as for which one fit the dataset better than the other techniques.

In chapter 1, the definitions and burden of peritonitis and peritoneal dialysis are explained. This was done as a way of breaking down the title of the study. The second chapter of the study focused on reviewing the work of other researchers in the context of peritonitis and survival analysis. In chapter 3, the statistical data analysis techniques which are applied to carry out the results were discussed. Finally, chapter 4 presented the results which were generated using the techniques discussed in chapter 3.

The univariate counting process, stratified, gap-time and marginal hazard regression models were applied to select the significant covariates to the multivariate regression hazard models. The multivariate regression model were compared as for which one fit the data better than the other models. The significant risk factors were identified from the better fitting model. Regression coefficient for covariates were considered to be statistically significant at 5% level.

The model results in this study are obtained using the Breslow method of handling ties, however, due to the small number of ties among the episodes times of the patients, the results are parallel to the ones obtained when utilising the Efron method of handling ties. The standard error of the regression coefficients were adjusted for the possible correlation among episode times coming from the same individuals through the sandwich robust variance estimator. The rate at which PD patients experiences recurrent episodes of peritonitis were measured using the hazard ratio. AIC and SBC were both employed to detect the best fitting recurrent survival model.

The study revealed that males, blacks and not married (single) individuals are the mostly followed patients in this study. This implies that kidney failure is a problem to

black men who are not married. However, they were found to be not the significant risk factors associated with recurrent episodes of peritonitis. The stratified proportional hazard model was the one having the smallest AIC and SBC values when compared to the other three modelling techniques. This was evident enough in the study to conclude that the stratified approach is the good modelling technique to fit the available recurrent events data set.

The stratified proportional hazard model illustrated that two categories of marital status, divorced and widower, and GFR-MDRD were the only factors found to be associated with recurrent episodes of peritonitis at 5% level of significance. However, the rate of experiencing recurrent episodes of peritonitis was high in widower and single patients as compared to the married patients. The study disclosed that an increase in GFR-MDRD and Hb lowers the rate of experiencing recurrent episodes of peritonitis. That is, PD patients with low GFR-MDRD and Hb are at high risk for recurrent episodes of peritonitis.

5.3 Limitations and recommendations

The Breslow and Efron partial likelihood techniques for ties handling works quite the same when the number of ties are small (Klein and Moeschberger, 2003). This correspond to the findings of this study. Therefore, the limitation of the study was that 152 PD patients were followed and they were not good enough to allow the comparison of Breslow and Efron partial likelihood techniques. The number of studies conducted on recurrent episodes of peritonitis are inadequate and therefore, this limit the study to have enough information to discuss in literature review. The analysis of the study was conducted using secondary data with few missing values and this was also considered as a study limitation because it is important to conduct a study with a complete information.

Duchateau et al. (2003) recommended the gap-time model as the most suitable approach for studying the recurrent events rate or multiple episodes data. However, this study through the AIC and SBC, recommended the stratified proportional hazard model as the best fitting technique for recurrent episodes data. Therefore, future researchers can utilise the stratified modelling technique when analysing recurrent events data set. Statisticians, health authorities and medical practitioners can use this information to educate people about recurrent episodes of peritonitis. This information can also be used to detect PD patients at high risk of experiencing recurrent episodes of peritonitis.

5.4 Areas for further study

There are numerous studies conducted for investigating the major risk factors associated with time-to-first episode of peritonitis. However, there are limited number of studies conducted on the major risk factors influencing the occurrence of more than one episode of peritonitis, that is, there are few studies available which focuses on the factors associated recurrent episodes of peritonitis. Since this study is one of the few studies on recurrent episodes of peritonitis, more studies should be conducted in the future in order to improve or add to the little discovered information.

This study aimed at utilising and comparing different recurrent survival analysis techniques to investigate the major risk factors associated with recurrent episodes of peritonitis. Therefore, it will be of great interest if future researchers can focus on recurrent episodes of peritonitis and improve the outcome of PD patients. Different modelling techniques such as, the frailty and AFT models, must be employed on investigating these risk factors associated with peritonitis.

5.5 Conclusion

This chapter presented the summary of the methods utilised to answer the aim and objectives of the study and have also summarised the finding related to the aim and objectives of the study. The findings of the study are important to the future researchers, statisticians and medical practitioners who are concerned with the health of the PD patients.

REFERENCES

BARONE, R., CÁMPORA, M., GIMENEZ, N., RAMIREZ, L., PANESE, S., AND SANTOPIETRO, M. (2012). Continuous ambulatory peritoneal dialysis versus automated peritoneal dialysis and peritonitis in the short and very long term at risk. *In Advances in peritoneal dialysis. Conference on Peritoneal Dialysis*, volume 28. pp. 44–49.

BORGAN, Ø. (1997). Three contributions to the encyclopedia of biostatistics: The nelson-aalen, kaplan-meier, and aalen-johansen. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*.

CHILDERS, D. D. (2015). *Summary of Survival Analysis with SAS Procedures*. Ph.D. thesis, University of Louisville.

CHOW, K. M., SZETO, C. C., LEUNG, C. B., KWAN, B., LAW, M. C., AND LI, P. (2005). A risk analysis of continuous ambulatory peritoneal dialysis-related peritonitis. *Peritoneal Dialysis International*, **25** (4), 374–379.

DUCHATEAU, L., JANSSEN, P., KEZIC, I., AND FORTPIED, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52** (3), 355–363.

FAN, X., HUANG, R., WANG, J., YE, H., GUO, Q., YI, C., LIN, J., ZHOU, Q., SHAO, F., YU, X., ET AL. (2014). Risk factors for the first episode of peritonitis in southern chinese continuous ambulatory peritoneal dialysis patients. *PloS one*, **9** (9), e107485.

FENG, S., WANG, Y., QIU, B., WANG, Z., JIANG, L., ZHAN, Z., JIANG, S., AND SHEN, H. (2016). Impact of early-onset peritonitis on mortality and technique survival in peritoneal dialysis patients. *SpringerPlus*, **5** (1), 1676.

FRIED, L. F., BERNARDINI, J., JOHNSTON, J. R., AND PIRAINO, B. (1996). Peritonitis influences mortality in peritoneal dialysis patients. *Journal of the American Society of Nephrology*, **7** (10), 2176–2182.

GARDINER, J. (2010). Survival analysis: overview of parametric, nonparametric and semiparametric approaches and new developments. *In SAS Global Forum 2010. Statistics and Data Analysis*.

GOLPER, T. A., BRIER, M. E., BUNKE, M., SCHREIBER, M. J., BARTLETT, D. K., HAMILTON, R. W., STRIFE, F., HAMBURGER, R. J., ET AL. (1996). Risk factors for peritonitis in long-term peritoneal dialysis: the network 9 peritonitis and catheter survival studies. *American journal of kidney diseases*, **28** (3), 428–436.

GRAY, N. A., GRACE, B. S., AND MCDONALD, S. P. (2013). Peritoneal dialysis in rural australia. *BMC nephrology*, **14** (1), 278.

HAN, S. H., LEE, S. C., AHN, S. V., LEE, J. E., KIM, D. K., LEE, T. H., MOON, S. J., KIM, B. S., KANG, S.-W., CHOI, K. H., ET AL. (2007). Reduced residual renal function is a risk of peritonitis in continuous ambulatory peritoneal dialysis patients. *Nephrology Dialysis Transplantation*, **22** (9), 2653–2658.

IKABU, A. S., ASSOUNGA, A. G. H., AND APALATA, T. (2007). A study of peritonitis in continuous ambulatory peritoneal dialysis patients at inkosi albert luthuli central hospital, durban, south africa.

ISLA, R. A. T., AMEH, O. I., MAPIYE, D., SWANEPOEL, C. R., BELLO, A. K., RATSELA, A. R., AND OKPECHI, I. G. (2016). Baseline predictors of mortality among predominantly rural-dwelling end-stage renal disease patients on chronic dialysis therapies in limpopo, south africa. *PloS one*, **11** (6), e0156642.

ISLA, R. A. T., MAPIYE, D., SWANEPOEL, C. R., ROZUMYK, N., HUBAHIB, J. E., AND OKPECHI, I. G. (2014). Continuous ambulatory peritoneal dialysis in limpopo province, south africa: predictors of patient and technique survival. *Peritoneal Dialysis International*, **34** (5), 518–525.

KELEŞ, M., CETINKAYA, R., UYANIK, A., ACEMOUGLU, H., SAATÇI, F., AND UYANIK, M. H. (2010). Peritoneal dialysis-related peritonitis: an analysis of risk factors in northeast anatolia. *Turkish Journal of Medical Sciences*, **40** (4), 643–650.

KLEIN, J. P. AND MOESCHBERGER, M. L. (2003). Techniques for censored and truncated data. statistics for biology and health.

KLEINBAUM, D. G. AND KLEIN, M. (2006). *Survival analysis: a self-learning text*. New York: Springer Science & Business Media.

KLEINBAUM, D. G. AND KLEIN, M. (2010). *Survival analysis*, volume 3. Springer.

KOTSANAS, D., POLKINGHORNE, K. R., KORMAN, T. M., ATKINS, R. C., AND BROWN, F. (2007). Risk factors for peritoneal dialysis-related peritonitis: Can we reduce the incidence and improve patient selection. *Nephrology*, **12** (3), 239–245.

LIU, X. (2012). *Survival analysis: models and applications*. John Wiley & Sons.

MARTIN, L. C., CARAMORI, J. C., FERNANDES, N., DIVINO-FILHO, J. C., PECOITS-FILHO, R., AND BARRETTI, P. (2011). Geographic and educational factors and risk of the first peritonitis episode in brazilian peritoneal dialysis study (brazpd) patients. *Clinical Journal of the American Society of Nephrology*, **6** (8), 1944–1951.

MARTINEZ, C. (2014). Uso del programa testsurvrec para comparar curvas de superviviencia con eventos recurrentes. *Revista Ingenieria UC*, **21** (2), 7–15.

MASHILOANE, B., MOSHESH, F., AND MPE, M. (2008). Peritonitis in patients with end-stage renal disease on continuous ambulatory peritoneal dialysis. *SAMJ: South African Medical Journal*, **98** (12), 942–944.

MEHROTRA, R., CHIU, Y.-W., KALANTAR-ZADEH, K., BARGMAN, J., AND VONESH, E. (2011). Similar outcomes with hemodialysis and peritoneal dialysis in patients with end-stage renal disease. *Archives of internal medicine*, **171** (2), 110–118.

NESSIM, S. J., BARGMAN, J. M., AUSTIN, P. C., NISENBAUM, R., AND JASSAL, S. V. (2009). Predictors of peritonitis in patients on peritoneal dialysis: results of a large, prospective canadian database. *Clinical Journal of the American Society of Nephrology*, **4** (7), 1195–1200.

NIETO-RÍOS, J. F., DÍAZ-BETANCUR, J. S., ARBELÁEZ-GÓMEZ, M., GARCÍA-GARCÍA, Á., RODELO-CEBALLOS, J., REINO-BUELVAS, A., SERNA-HIGUITA, L. M., AND HENAO-SIERRA, J. E. (2014). Peritoneal dialysis-related peritonitis: twenty-seven years of experience in a colombian medical center. *Nefrologia*, **34** (1), 88–95.

OKAYAMA, M., INOUE, T., NODAIRA, Y., KIMURA, Y., NOBE, K., SETO, T., SUEYOSHI, K., TAKANE, H., TAKENAKA, T., OKADA, H., ET AL. (2012). Aging is an important risk factor for peritoneal dialysis-associated peritonitis. *In Advances in peritoneal dialysis. Conference on Peritoneal Dialysis*, volume 28. pp. 50–54.

OO, T. N., ROBERTS, T. L., AND COLLINS, A. J. (2005). A comparison of peritonitis rates from the united states renal data system database: Capd versus continuous cycling peritoneal dialysis patients. *American journal of kidney diseases*, **45** (2), 372–380.

PORT, F. K., HELD, P. J., NOLPH, K. D., TURENNE, M. N., AND WOLFE, R. A. (1992). Risk of peritonitis and technique failure by capd connection technique: a national study. *Kidney international*, **42** (4), 967–974.

PRENTICE, R. L., WILLIAMS, B. J., AND PETERSON, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 373–379.

RICHÉ, F. C., DRAY, X., LAISNÉ, M.-J., MATÉO, J., RASKINE, L., SANSON-LE PORS, M.-J., PAYEN, D., VALLEUR, P., AND CHOLLEY, B. P. (2009). Factors associated with septic shock and mortality in generalized peritonitis: comparison between community-acquired and postoperative peritonitis. *Critical Care*, **13** (3), R99.

ROBERTO, D., MARIA, P. R., EMILIA, S., PIERLUIGI, D. L., CATALINA, O. K., DINNA, C. N., POLANCO, D. K., , VALENTINA, C., CAL, M. D., AND RONCO, C. (2007). Advances in the technology of automated, tidal, and continuous flow peritoneal dialysis. *PloS one*, **27** (2), 130 – 137.

RUDNICKI, M., KERSCHBAUM, J., HAUSDORFER, J., MAYER, G., AND KÖNIG, P. (2010). Risk factors for peritoneal dialysis–associated peritonitis: the role of oral active vitamin d. *Peritoneal Dialysis International*, **30** (5), 541–548.

SCHNEIDER, C. P., SEYBOTH, C., VILSMAIER, M., KÜCHENHOFF, H., HOFNER, B., JAUCH, K.-W., AND HARTL, W. H. (2009). Prognostic factors in critically ill patients suffering from secondary peritonitis: a retrospective, observational, survival time analysis. *World journal of surgery*, **33** (1), 34.

STEVENSON, M. AND EPICENTRE, I. (2009). An introduction to survival analysis. *EpiCentre, IVABS, Massey University*.

VONESH, E. F. (1985). Estimating rates of recurrent peritonitis for patients on capd. *Peritoneal Dialysis International*, **5** (1), 59–65.

ZENT, R., MYERS, J., DONALD, D., AND RAYNER, B. (1994). Continuous ambulatory peritoneal dialysis: an option in the developing world. *Peritoneal Dialysis International*, **14** (1), 48–51.