Identification of factors affecting the survival lifetime of HIV+ terminal patients in Albert Luthuli municipality of South Africa

by

PEPUKAI BENGURA

submitted in accordance with the requirements
for the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: Prof P NDLOVU

CO-SUPERVISOR: Ms M A MANAGA

19 December 2019

ABSTRACT

The objective of the study was to identify the factors that affect the survival lifetime of HIV+ terminal patients in rural district hospitals of Albert Luthuli municipality in the Mpumalanga province of South Africa. A cohort of HIV+ terminal patients was retrospectively followed from 2010 to 2017 until a patient died, was lost to follow-up or was still alive at the end of the observation period. Nonparametric survival analysis and semiparametric survival analysis methods were used to analyse the data. Through Cox proportional hazards regression modelling, it was found that ART adherence (poor, fair, good), Age, Follow-up mass, Baseline sodium, Baseline viral load, Follow CD4 count by Treatment (Regimen 1) interaction and Follow-up lymphocyte by TB history (yes, no) interaction had significant effects on survival lifetime of HIV+ terminal patients (p-values<0.1). Furthermore, through quantile regression modelling, it was found that short, medium and long survival times of HIV+ patients, respectively represented by the 0.1, 0.5 and 0.9 quantiles, were not necessarily significantly affected by the same factors.

Keywords:

Survival time, Follow-up time, HIV/AIDS disease, Antiretroviral therapy, ART adherence, CD4 cell count, Cox proportional hazards regression, Logistic regression, Quantile regression, Kaplan-Meier estimator, Log-rank test.

DECLARATION

I declare that this thesis is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Identification of factors affecting the survival lifetime of HIV+ terminal patients in Albert Luthuli municipality of South Africa.

Declared on the (date): 19 December 2019

Signed: 

Name: Pepukai Bengura

Declared on the (date): 19 December 2019

Signed: 

Name: Prof P. Ndlovu (Supervisor)

Signed: 

Name: Ms M.A Managa (Co-supervisor)

DEDICATION

I dedicate this work to my late father S.S Bengura who sowed a seed of education in my life. My mothers too, played exceptional contributory roles in shaping me. Above all, I give all the glory and honour to the Almighty God in whom all the faculties of my life are nested.

ACKNOWLEDGEMENTS

Supervisor Prof P. Ndlovu and Co-Supervisor A. M. Managa assisted me in shaping the thesis title, in academic writing, in thesis layout and in statistical analyses using different multivariable regression methods and survival techniques. The supervisors were my academic compass throughout the winding and undulating research journey. They never left me alone in academic doldrum.

Data Capturers at Embhuleni and Carolina hospitals played a pivotal role in tiresome retrieval of files; a process which accelerated my research. They also gave some moral support. The Management of Embhuleni and Carolina hospitals played professional roles which made it possible for the research to take off. Thereafter, they remained supportive.

UNISA Ethics Review Committee and Mpumalanga Department of Health need special applaud for their speedy processing of my applications regarding ethics.

My wife Patricia was supportive in many ways until the final day. My children; namely: Yolanda, Ashton, Lerato and Andile were tolerant to my absence from home as I was pursuing my studies.

The list is tall, and to those not mentioned; I still remember you and I will forever value the invaluable contributions you made.

ABBREVIATIONS, ACRONYMS AND DEFINITIONS

3TC	Lamivudine
ABC	Abacavir
AIDS	Acquired immunodeficiency Syndrome
AFT	Accelerated Failure Time
ALT	Alanine amino transferase
ART	Antiretroviral therapy
ARV	Antiretroviral
AZT	Zidovudine
ABC	Abacavir
BMI	Body Mass Index
CD4	Cluster of Differentiation 4, a glycoprotein that is found primarily on the surface of helper T cells
CI	Confidence interval
CQR	Censored Quantile Regression
d4T	Stavudine
Ddl	Didanosine
DHIS	District Health Information System
DoH	Department of Health
EFV	Efavirenz
HAART	Highly Active Antiretroviral Therapy
Hb	Hemoglobin
HBV	Hepatitis B virus
HR	Hazard ratio
MDoH	Mpumalanga Department of Health
NFV	Nelfinavir
NGO	Non-Governmental Organisation
NVP	Nevirapine
R	Freely available statistical analysis software
TB	Tuberculosis
SAS	Statistical Analysis System
SPSS	Statistical Package for Social Science
Stata	Statistical software package created by Stata Corp
TB	Tuberculosis
UNAIDS	United Nations Programme on HIV/AIDS
USAID	United States Agency for International Development
WBC	White blood count
WHO	World Health Organisation

Main Source for most abbreviations above: Ministry of Health and Social Services, Directorate of Special Programmes. (2010). Guideline for anti-retroviral therapy. Windhoek, Namibia.

DEFINITIONS OF MAIN TERMS

- HIV+ terminal patient -patient experiencing progressive or worsening HIV/AIDS and who is going to die due to HIV/AIDS or related health problems.
- Transferred patient - a patient who leaves a current health facility to another health facility because of personal reasons or because of HIV treatment failure which may include drug resistance, drug toxicity or poor adherence to antiretroviral therapy (ART).
- Lost to follow-up patient - a patient who gives a threat to the long-term success of antiretroviral therapy programme by failing to visit the clinic for 90 days or longer from the last attendance for refill and not yet classified as ‘dead’ or ‘transferred-out’.
- Died patient – a patient who dies under the administration of an antiretroviral therapy programme as a result of HIV/AIDS or associated factors.

TABLE OF CONTENT

CHAPTER 1 ..... **1**
INTRODUCTION 1
1.1 Background of the study 1
1.2 Previous studies on survival of HIV/AIDS patients 3
1.3 Problem statement and research question 5
1.4 Purpose of the study 5
1.5 Significance of the study 6
1.6 Data 7
1.6.1 Study area 7
1.6.2 Study design 7
1.6.3 Data sources and cleaning of data 8
1.6.4 Sampling procedures and demographics 8
1.6.5 Variables in the data set 11
1.7 Limitations and delimitations of the study 13
1.8 Assumptions 13
1.9 Ethical considerations 14
1.10 Summary 14
1.11 Organisation of the study 14
CHAPTER 2 ..... **16**
REVIEW OF SURVIVAL ANALYSIS METHODS 16
2.1 Introduction 16
2.2 Nonparametric survival analysis methods 18
2.2.1 The Kaplan-Meier estimator of the survivor function 18
2.2.2 Comparison of survivor functions 19
2.3 Semiparametric survival analysis methods 22
2.3.1 The Cox proportional hazards (PH) model 22
2.3.1.1 The theoretical aspects of the Cox PH model 22
2.3.1.2 The application of the Cox model 23
2.3.2 Inferences about the Cox model parameters 23
2.3.3 Model selection and diagnostics in Cox PH models 25
2.3.3.1 Model and variable selection 25
2.3.3.2 Proportional hazards and linearity of the predictor assumptions 27
2.3.3.3 Outliers and influential observations 28
2.3.3.4 Goodness of fit 30

2.3.4 Interpretation of the coefficients of the Cox PH model 31
2.4 Logistic regression in survival analysis 32
2.4.1 Inferences about the logistic regression model parameters 33
2.4.2 Model selection and diagnostics in logistic regression 34
2.4.2.1 Model and variable selection 34
2.4.2.2 Outliers and influential observations 35
2.4.2.3 Appropriateness of the linearity of predictor 35
2.4.2.4 Goodness of fit 35
2.4.5 Interpreting the fitted logistic regression model 37
2.5 Quantile regression in survival analysis 38
2.5.1 Inferences about the quantile regression model parameters 39
2.5.2 Model selection and diagnosis in quantile regression models 41
2.5.3 Interpretation of the coefficients 43
2.6 Summary 44
CHAPTER 3 45
DATA ANALYSIS AND RESULTS ..... ….45
INTRODUCTION 45
3.1 Preliminary data analysis 45
3.2 Semiparametric and nonparametric analysis of the data 51
3.2.1 Cox PH regression modelling 51
3.2.1.1 Checks for outliers and influential observations 52
3.2.1.2 Checks of the linearity and proportional hazards assumptions 56
3.2.1.3 Checking for the overall goodness of fit of the model 57
3.2.1.4 Harrel's concordance statistic 58
3.2.1.5 Estimated parameters of the final Cox PH model 58
3.2.2 Nonparametric inferences about the survivor functions 62
3.2.2.1 Kaplan-Meier survival functions for ART adherence groups 62
3.2.2.2 Kaplan-Meier survival functions for Age groups 63
3.2.2.3. Kaplan-Meier survival functions for Follow-up mass 64
3.2.2.4. Kaplan-Meier survival functions for Baseline sodium 64
3.2.2.5. Kaplan-Meier survival functions for Baseline viral load 65
3.2.2.6 Kaplan-Meier survival functions for Follow-up lymphocyte by TB history groups 66
3.2.2.7: Kaplan-Meier survival function estimates for Follow-up CD4 by Treatment (Regimen 1) 67
3.2.2.8 Kaplan-Meier survival functions for Follow-up CD4 and related results 68

3.3 Quantile regression modelling 71
3.3.1 Quantile regression modelling at quantile levels 0.1, 0.5 and 0.9 71
3.3.2 Some quantile process plots to study heterogeneity in the data 76
3.3.3 Survival analysis of the effect of covariates on quantiles of the survival time of patients 78
3.4 Summary 79
CHAPTER 4 80
FINDINGS AND DISCUSSIONS 80
4.1 Findings 80
4.1.1 Findings based on demographic and outcome factors 80
4.1.2 Findings based on Logistic regression 81
4.1.3 Findings based on Cox regression 81
4.1.4 Findings based Kaplan-Meier survival functions and Hazard ratios 82
4.1.4.1 ART adherence 82
4.1.4.2 Age 83
4.1.4.3 Follow-up mass 83
4.1.4.4 Baseline viral load 83
4.1.4.5 Follow-up lymphocyte by TB history groups 84
4.1.4.6 Follow-up CD4 by Treatment (Regimen 1) 84
4.1.4.7 Follow-up sodium 85
4.1.5 Findings based on Quantile regression 86
4.1.6 Findings based on SAS output for 95% hazard ratio confidence intervals for small samples and heavy censoring 87
4.2 Discussions 87
CHAPTER 5 94
CONCLUSION AND RECOMMENDATIONS 94
5.1 Conclusion 94
5.2 Recommendations 95
References 97

TABLES

Table 1.6.1: Distribution of patients by Hospital, Gender and Age in the study population 9
Table 1.6.2: Distribution of dead patients over Follow-up time (in years) 10
Table 1.6.3: Distribution of dead and censored patients by Gender, Age group and Hospital in the sample 10
Table 1.6.4: Distribution of dead patients by ART adherence, TB history, WHO Stage and Treatment (Regimen 1) in the sample 11
Table 1.6.5: List of variables in the data set 12
Table 3.1.1: Frequency distribution of Patient status by Hospital on last follow-up visit 45
Table 3.1.2: Frequency distribution of Patient status by Gender on last follow-up visit 46
Table 3.1.3: Frequency distribution of the Patient status by Age group on last follow-up visit 47
Table 3.1.4: Results of fitting the logistic regression model with binary response Patient status (0=Transferred out of hospital or Lost to follow-up or Still alive and 1=Dead) 48
Table 3.1.5: Results of fitting the logistic regression model with binary response Patient status (0=Dead or Still alive and 1= Transferred out of hospital or Lost to follow-up) 50
Table 3.2.1: Results of fitting the Cox PH regression model (right censored times are for Transferred out of hospital or Lost to follow-up or Still alive patients) 52
Table 3.2.2: A typical table for checking the presence of outliers in a Categorical covariate 53
Table 3.2.3: Influential observations with respect to more than one covariate in Table 3.2.1 55
Table 3.2.4: The results of the Supremum test for the linearity of the predictor assumption 56
Table 3.2.5: The results of the Supremum test for proportional hazards assumption 57
Table 3.2.6: Estimated Harrel's concordance statistic 58
Table 3.2.7: Maximum likelihood estimates of the final Cox PH model 61
Table 3.2.8: Log-rank and other tests for equality of ART adherence survivor functions 63
Table 3.2.9: Log-rank and other tests for equality of Age group survivor functions 63
Table 3.2.10: Log-rank and other tests t for equality of Follow-up mass survivor functions 64
Table 3.2.11: Log-rank and other tests for equality of Baseline sodium survivor functions 65
Table 3.2.12: Log-rank and other tests for equality of Baseline viral load survivor functions 66
Table 3.2.13: Log-rank and other tests for equality of Follow-up lymphocyte by TB history survivor functions 67
Table 3.2.14: Log-rank and other tests for equality of Follow-up CD4 count by Treatment (Regimen 1) survivor functions 67
Table 3.2.15: Log-rank and other tests for equality of Follow-up CD4 strata survivor functions 68
70
Table 3.2.16: Hazard ratios for the main and interaction effects strata 70
Table 3.3.1: Results of fitting the quantile regression model at quantile level 0.1 73
Table 3.3.1: Results of fitting the quantile regression model at quantile level 0.1 73
Table 3.3.2: Results of fitting the quantile regression model at quantile level 0.5 74
Table 3.3.2: Results of fitting the quantile regression model at quantile level 0.5 74
Table 3.3.3: Results of fitting the quantile regression model at quantile level 0.9 75
Table 3.3.3: Results of fitting the quantile regression model at quantile level 0.9 75
Table 3.3.4: Likelihood Ratio and Wald Tests of significance of the quantile models in the tables 3.3.1, 3.3.2 and 3.3.3 76
Table 3.3.4: Likelihood Ratio and Wald Tests of significance of the quantile models in the tables 3.3.1, 3.3.2 and 3.3.3 76

FIGURES

Figure 3.2.1: A typical Box plot for checking the presence of outliers in the values of a covariate 53
Figure 3.2.2: The DFBETAS graph for covariate Follow-up CD4 in the model in Table 3.2.1 54
Figure 3.2.3: The DFFITS graph for checking the presence of influential observations. 55
Figure 3.2.4: Estimated Nelson-Aalen cumulative hazard function versus the Cox-Snell hazard residuals (dotted line): continuous 45° theoretical line 58
Figure 3.2.5: Kaplan-Meier survival function estimates for ART adherence strata 63
Figure 3.2.6: Kaplan-Meier survival function estimates for Age groups 63
Figure 3.2.7: Kaplan-Meier survival function estimates for Follow-up mass strata 64
Figure 3.2.8: Kaplan-Meier survival function estimates for Baseline sodium strata 65
Figure 3.2.9: Kaplan-Meier survival function estimates for Baseline viral load strata 66
Figure 3.2.10: Kaplan-Meier survival function estimates for Follow-up lymphocyte by TB history strata 67
Figure 3.2.11: Kaplan-Meier survival function estimates for Follow-up CD4 by Treatment (Regimen 1) strata 67
Figure 3.2.12: Kaplan-Meier survival function estimates for Follow-up CD4 Count 68
Figure 3.2.13: Bar graphs for the distribution of Baseline and Follow-up CD4 in HIV/AIDS patients 69
Figure 3.3.1: Estimated parameter versus quantile for log (survival time) with 95% confidence limits - Follow-up CD4, *ln (*Baseline viral load) and intercept 76
Figure 3.3.2: Estimated parameter by quantile for *ln* (survival time) with 95% confidence limits - ART adherence and Treatment (Regimen 1) 77
Figure 3.3.3: Survival analysis of a continuous covariate effect on quantiles of Survival time 78
Figure 3.3.4: Survival analysis of a categorical covariate effect on quantiles of Survival time 79

APPENDICES

Appendix A: Research Data capturing tool (Excel) xcvi
Appendix B: Research Data storage tool (Excel) xcvi
APPENDIX C: Summary of the big picture of AIDS xcvi
APPENDIX D: South Africa in first position of people living with HIV xcvi
Appendix E: Map showing the relative position of Chief Albert Luthuli in Gert Sibande District xcvi
Appendix F: Map showing the distribution of the Health sites in Albert Luthuli Municipality xcvi
Appendix G: Important statistics on Albert Luthuli Municipality. xcvi
Appendix H: Descriptions of technical independent variables part 1 xcvi
Appendix I: Descriptions of technical independent variables part 2 xcvi
Appendix J: Laboratory Report on Sodium, Creatinine, Kidney damage (eGFR) Total protein and ALT xcvi
Appendix K: Laboratory Report on Haemoglobin xcvi
Appendix L: Laboratory Report on Viral Load xcvi
Appendix M: Laboratory Report on CD4 count, Lymphocyte and White blood cell count xcvi
Appendix N: Mpumalanga Department of Health Permission letter xcvi
Appendix O: UNISA Ethics Approval xcvi
Appendix P: Main statistical programming used: SAS, R and Stata xcvi

CHAPTER 1

INTRODUCTION

This chapter describes the background of Human Immuno-Deficiency Virus/Acquired Immune Deficiency Syndrome (HIV/AIDS) with South Africa as the point of focus. Previous studies on survival analysis of HIV/AIDS patients, the research problem, the research question, the purpose of the study and the significance of the study are the components of this chapter. The Data section in this chapter has the following subheadings: study area, study design, data source and data cleaning, sampling procedures and demographics and variables in the data. The study limitations and delimitations, assumptions of the study and ethical considerations are also included in this chapter.

1.1 Background of the study

HIV/AIDS has been a major health problem worldwide for more than three decades now. The 'big picture' of AIDS is shown in Appendix C. Sub-Saharan Africa has the worst HIV and AIDS epidemic in the world as indicated in Appendix D. According to United Nations Agency for International Development (UNAIDS) Gap Report (2016), South Africa has the biggest HIV epidemic in the world, with estimated 7 million (12.7%) living with HIV in 2015. In the same year, there were 380 000 (0.7%) new infections while 180 000 (0.3%) South Africans died from AIDS-related illnesses (UNAIDS Gap Report, 2016). Statistics South Africa Report (2015) indicates a rise in AIDS-related deaths from 29.2% in 2014 to 30.5% in 2015.

South Africa's Mpumalanga province has the second highest HIV prevalence rate after KwaZulu-Natal province, and the Gert Sibande district which is in Mpumalanga is leading all districts in the country with 46.1% prevalence rate (Masinga, 2014; Motsoaledi, 2013). The Gert Sibande district HIV prevalence stood at 40.5% in 2011 (Bezuidenhout et al., 2014) and at 38.6% in 2015 (National Department of Health, 2015). The Gert Sibande district has Albert Luthuli as one of its seven municipalities (Appendix E). In comparison to Gert Sibande, the HIV prevalence for Albert Luthuli stood at 43.2% in 2011 Nkosi (2017), and the prevalence stood at 44% in 2014 (Chief Albert Luthuli Municipality, 2018). The following, which could be pointers to HIV/AIDS prevalence and effects, are own current observations in Albert Luthuli municipality.

- Comparatively high number of pregnant girls in high schools. In Gert Sibande in 2016/17, there were 1 399 teenage pregnant girls, and the number increased by about 85.9% to 2 601 in 2018 (Mpande, 2018). A teenage pregnancy rise of 78% was reported in Mpumalanga in just one year ("Nearly 6 000 babies born," 2018).
- Significantly large number of child-headed families. Of all 2088 child-headed households in Gert Sibande district, 534 (about one-quarter) of them are in Albert Luthuli municipality (Electoral Commission of South Africa [IEC], 2016).
- Behaviours and activities that significantly point to promiscuous sexual activities in some communities.
- Articles which cited Gert Sibande as one of the districts in South Africa which is hard hit by HIV/AIDS.
- The proximity of Albert Luthuli municipality to Eswatini. Eswatini, is country in southern Africa, and has the highest HIV prevalence in the world, with 27.3% of adults living with HIV (Avert, 2017).

South Africa is known to be having the largest antiretroviral treatment (ART) programme globally. The number of patients in Gert Sibande receiving ART up to November 2011 was 32 979 (3.2%) (Bezuidenhout et al., 2014), and this number rose to 76 632 (6.75%) in 2015 (Mpumalanga Provincial AIDS Council, 2016). South Africa ART programme has saved millions of lives and infections averted. However, the dual burden of Tuberculosis (TB) and HIV disease continues unabated. South Africa continues to see an increase in HIV related TB incidences and has not yet felt the impact of HIV prevention programmes. There are some statistical methods such as survival analysis methods that can be used to analyse data for monitoring the survival time benefit of HIV/AIDS interventions, and for investigating potential factors that may affect the survival probability of HIV/AIDS patients.

The theories of survival analysis methods are reviewed in (chapter 2). Briefly, in the context of this thesis, survival analysis methods are used to estimate the risk of death or progression of a disease and to provide predictions that help clinicians to estimate trends in their patient outcomes (Nakhae & Law, 2011). On one hand, the methods are used to estimate the time period during which an event (for example, death) can happen, and on the other hand, to estimate the impact of various independent factors on the time distribution to the occurrence of an event (Melnyk et al.,1995).

1.2 Previous studies on survival of HIV/AIDS patients

There is no evidence that survival analysis methods have ever been fully used to analyse survival lifetime patterns of HIV/AIDS patients in Albert Luthuli municipality. This, and the aforementioned about the HIV/AIDS disease in Albert Luthuli motivated this study. In previous studies on survival analysis, Kaplan-Meier estimator was used to estimate mortality (Etikan et al., 2017), and according to Cox (1972), Tadesse et al. (2014), Walters (2012) and Wilson (2013), Cox model was used to identify predictors of mortality.

Damtew et al. (2015) carried out a retrospective cohort study to identify predictors of survival among 784 HIV-infected adult patients on ART at a public hospital in eastern Ethiopia. The study by Damtew et al. (2015) had estimated mortality rates as 8.4%, 9.8%, 11.3%, 12.7% and 14.1% at 6, 12, 24, 36 and 48 months, respectively. The predictors with significant values of mortality in this study were marital status (high for single), functional status (high for bedridden), World Health Organisation (WHO) stage (high for stage 4), Body Mass Index (BMI) (high for big deviation from the normal), Cluster of Differentiation 4 (CD4) count (high for low values) and anaemia (high for low values). Improved survival probability was observed in patients taking ART.

In a related study in Ethiopia, Moshago et al. (2014) researched on the survival and risk factors of mortality in people living with HIV/AIDS. The Cox proportional hazard regression model was used for estimating the survival time to death. The cumulative survival probabilities of patients at the 6^{th} month after initiation of ART were 96%, 94% and 96% for children, adolescents and adults, respectively. The mortality rate was significantly higher among the patients with the low CD4 count, advanced WHO stage, isonicotinic acid hydrazide (INH) prophylaxis, TB infection and bedridden functional status. These researchers concluded low mortality rate and a high rate of the loss to follow-up of the cohort. They also concluded that a higher proportion of adolescents was lost to follow-up in comparison to child and adult age groups.

In Far-North Province of Cameroon, Sieleunou et al. (2008) analysed the outcomes of ART from a retrospective cohort study of 1187 rural patients with ages greater than 15 years using the Kaplan-Meier estimator and the Cox model. The follow-up was from 2001 to 2006. The survival probability was 77% at 1 year and 47% at 5 years. The predictors of mortality were

CD4 count, haemoglobin, BMI, sex and clinical stage at enrolment. The researchers reported unique arrangements on counselling and on follow-up of those who dropped out. On adherence to counselling, every patient was supposed to bring along a relative who was trained to support the patient in adhering to the treatment. Non-adherent patients were referred to special counselling sessions and defaulters were actively traced by a messenger or by telephone. The patients who failed to return for follow-up visits for more than 3 months were classified as 'lost to follow-up' on condition that they were still alive and did not transfer out. Rigorous tracing of the patients lost to follow-up was done by the ART centre. A patient lost to follow-up who happened to die at home, was categorised in the data analysis as having died immediately after the last registered follow-up visit.

In South Africa, some similar studies were carried out in Tshwane district by Mlangeni et al. (2016), and in Johannesburg by (Sanne et al., 2009). Mlangeni et al. (2016) analysed the outcomes of ART from a cohort of 381 HIV patients who were retrospectively followed over 5 years from 2007 to 2011. The main statistical method used for data analysis in this study was logistic regression. The overall mortality rate was 5% after 5 years and the rate of long term-term retention in care was 57.4% (calculated excluding died patients, patients lost to follow-up and patients transferred to another site). After 6 months on Highly Active Antiretroviral Therapy (HAART), the mean rise in CD4 count was 113 cells/mm^3, and after 60 months it was 288 cells/mm^3. Viral load suppression to less than 400 copies/ mm^3 was achieved in 74% of the patients at 6 months and in 91% of the patients in 60 months. The significant predictors of survival in Tshwane district study were CD4 count and viral load. Sanne et al. (2009) retrospectively followed 7583 HIV-infected patients from Johannesburg area for four years from 2004 to 2007. The Kaplan-Meier estimator and the Cox model were the main methods used for statistical analysis. The overall mortality rate was low (2.9 deaths per 100 person-years or 0.2% over 4 years) and the rate of long-term retention in care was relatively high, it was 74.4% at the end of 4 years. In the 4^{th} year, the majority (59.6%) of patients had at least first line drug (mainly stavudine) substituted. In comparison to men, women were twice as likely to experience drug substitution. At 6 months, 90.8% of patients had suppressed virus to a viral load below 400 copies/ml. Only 10% females and 13% males initiated second line HAART. The predictors of survival in the Johannesburg study were retention in care, ART regimens, CD4 count and viral load. It is against most of these findings that the study is going to be based.

1.3 Problem statement and research question

Despite an array of expensive treatments and preventive interventions used to combat HIV/AIDS in South Africa, the prevalence and death toll due to the disease, still remain too high as was mentioned in (section 1.1 and 1.2). Consequently, the broad question is "Are there other factors, that include treatments and preventive interventions, which negatively or positively significantly affect the survival lifetime of HIV/AIDS patients in South Africa despite the availability of ARTs?". Potential such factors which are reported in literature, which may or may not apply to the South African environment, have been mentioned in (section 1.2). Other factors to be investigated are listed in Table 1.6.5 (section 1.6). In this study the question is narrowed down to Albert Luthuli municipality with the hope that the answer from this study is close to the answer of the broad question.

1.4 Purpose of the study

The development of most survival analysis methods has been driven by the biomedical need to improve the health of patients (Li & Ma, 2013). The main objective of this study is to use survival analysis methods to identify factors that affect the survival lifetime of HIV/AIDS patients in South Africa (despite the availability of ARTs), using data from Albert Luthuli municipality hospitals. The aim of the study concurs with the policy of the DOH of South Africa which is 'Working towards long healthy life to South Africans'.

The uniqueness of this study points to some of the study purposes as listed below.

- Most of the similar studies in South Africa, for example Mlangeni and Senkubuge (2016) and Sanne et al. (2009) have been limited to a follow-up period of 4-5 years. In this study the follow-up period is 7.5 years which, hopefully, improves the quality of the findings.
- Most of the similar studies carried in South Africa and other countries have concentrated mostly on patients in urban hospitals for example Sanne et al. (2009) in Johannesburg, and Desai et al. (2015) in Surat City of India. This study focused on patients in marginalised rural area hospitals which presents its own peculiar challenges.
- Most of the similar studies as in the case of Damtew et al. (2015) and Moshago et al. (2014) focused mainly on death and loss to follow-up, this study further checked the impact of most used ART regimens on the kidneys, liver and on patient mass anomalies (deviation from the normal mass). In view of this gap, initiation of ART drugs among HIV+ patients

may be likened to a tropical storm which brings life to a drought-stricken area, but at the same time leaving some pockets of destruction.

- The usual mean regression survival analysis method is supplemented with quantile regression and logistic regression survival analysis methods. Quantile regression is supposed to obtain findings that are robust to the violation of the assumption of the mean regression survival analysis models. Furthermore, the quantile regression approach is more informative than the mean regression in the sense that it allows the investigation of the factors that affect quantile survival times of patients whose results are or maybe of interest. Logistic regression complements the Cox regression method by specifically identifying the factors affecting the mortality of HIV+ terminal patients.

1.5 Significance of the study

This study intends to bring in some new body of knowledge by closing a gap or gaps as outlined above. Despite South Africa being one of the countries in Africa that is hard hit by HIV/AIDS, information on survival analysis of HIV/AIDS patients from Africa is limited and comes from studies of short duration with relatively high loss to follow-up (Abebe, 2014). Hence, the need and significance of this study.

The significance of this study also lies in its uniqueness as was mentioned above. Regression (mean/quantile) survival probability modelling allows survival probabilities to be predicted beyond the follow-up period, and thus allowing HIV/AIDS health planners to have well-informed predictions on ART administration and its associated outcomes on HIV/AIDS patients. Therefore, this study is a sharp tool for accountability and auditing of the scaled-up HIV/AIDS programmes. Furthermore, from case studies in Ethiopia, Brazil, Australia, India, Europe and North America and United Kingdom carried out by Moshago et al. (2014), Luz et al. (2016), McManus et al. (2012), Desai et al. (2015), Trickey et al. (2017) and by Croxford et al. (2017) respectively, predicted deaths agreed with the observed deaths following HIV/AIDS infection.

The application of the findings of this study in Albert Luthuli municipality could see the management of HIV/AIDS patients done in a different and optimistic way. According to Alemu and Sebastian (2010), a few studies have been conducted to investigate the survival benefit of ART scale-up programmes in Africa.

Lastly, the impact of HIV/AIDS in this study concurs with Karim et al. (2010) who view AIDS as incapacitating families and communities, stressing health care services and decimating students and teachers in schools. Business has not been spared, it continues to suffer losses of personnel productivity; economic growth is being undermined and scarce resources must be diverted to deal with the consequences of the epidemic. Therefore, this research is a significant step in addressing a diversity of long-standing challenges posed by HIV/AIDS.

1.6 Data

1.6.1 Study area

Albert Luthuli local municipality is situated in the Gert Sibande district of Mpumalanga as shown in Appendix E. Economic activities that are dominant in the municipality include agriculture, forestry, tourism, manufacturing and mining. The total area of the municipality is 5559km^2, and the population density was 33.8 people per km^2 in 2016. Some other important statistics of the municipality are summarised in Appendix G.

Carolina and Embhuleni district hospitals, from which data was collected, render comprehensive health care service which includes HIV and TB related treatments, care and support services to the surrounding communities in Albert Luthuli municipality. These hospitals are accredited antiretroviral (ARV) treatment initiation and on-going treatment sites. Embhuleni hospital is a 179-bed hospital, while Carolina is a 79-bed hospital (Chu et al., 2011). Carolina hospital is situated 79km to the West of Embhuleni hospital. Both hospitals serve mostly the rural population in Albert Luthuli municipality. The locations of the hospitals in Albert Luthuli local municipality are shown in the map in Appendix F. The two hospitals routinely record HIV+ patient information shown in record forms in Appendices A and B. The information is collected for the purpose of clinical monitoring and evaluation of patients. These patient records were raw secondary data for this study.

1.6.2 Study design

This study has a typical retrospective cohort longitudinal design. The follow-up start time for all patients is randomly distributed from 01/01/2010 to 31/03/2010. The maximum length of the follow-up time is 7.5 years with the closure of the observation window on 30 June 2017. The follow-up time for each patient started at the time the patient got initiated to the ART

programme at the hospital's wellness centre. The cohort of HIV+ terminal patients on ART was followed until a patient died or until a patient was censored as a result of loss to follow-up, transferred to another hospital or as a result of the conclusion of the study. The cohort of patients was divided into the following groups: children (age ≤ 10 years), adolescents (age 11-19 years) and adults (age ≥ 20 years) as was done in previous studies by Bakanda et al. (2011) and by Moshago et al. (2014). Each stratum was further divided into males and females. The main event of interest in this research was death of an HIV+ terminal patient. The patients who transferred out of the hospital or got lost to follow-up were classified as drop-out patients. The drop-out patients as stated in the assumptions on survival analysis in chapter 2, are assumed to have the same survival prospects as those who continue to be followed and still routinely reporting to a hospital until the conclusion of the study.

1.6.3 Data sources and cleaning of data

Data for the study was obtained from each hospital's electronic SOZO (electronic patient management system) database and from the standard patient files. The data of the study variables were captured in excel (appendices A and B). Some common challenges which arise in the analysis of survival data are due to missing data, influential observations, and existence of variable outliers in the data. Methods of dealing with these challenges are reviewed in chapter 2. Graphical methods based on the analysis of martingale, score and deviance residuals are used in identification of outliers as proposed by Therneau et al. (1990). The 'DFBETAS' and the 'DFFITS' methods as discussed in chapter 2, are used in this study for identification of outliers and influential observations as in Canette (2014), Cleves et al. (2010) and Mack (2016).

1.6.4 Sampling procedures and demographics

This study includes all HIV+ terminal patients who were admitted and started ART treatment in Carolina and Embhuleni hospitals regardless of age. The HIV+ terminal patients were excluded from the study upon the researcher's discretion. However, missing data on the following clinically significant variables guided the exclusion: ART initiation date, survival status, survival status date and ART regimen.

An optimum sample size was determined using Monkey Survey online sample size calculator and also using Rao soft online sample size calculator (Rao & Rao, 2004), with a margin of

error of 5% or 95% confidence level. The sample size needed for the study and matching the resource constraints and population proportions is 357. A stratified random sampling method was used to select the study sample. The standard patient registration file numbers (folder numbers) from each hospital population were sequentially listed in excel to form gender strata. Excel generated random numbers were assigned to each file number. The random numbers assigned to the file numbers were then sorted in ascending order and then simple random method was applied according to D'Auria (2013) on each gender per hospital to get the distribution of patients by hospital, by gender and by age as shown in Table 1.6.1.

Table 1.6.1: Distribution of patients by Hospital, Gender and Age in the study population

Chief Albert Luthuli Municipality	Patients initiated on ART from 01/01/2010 to 31/03/2010 at Carolina and Embhuleni Hospitals [719]											
Hospital population size	Carolina Hospital 83 (11.5%)						Embhuleni Hospital 636 (88.5%)					
Hospital population size by gender	Females 50 (60.2%)			Males 33 (39.8%)			Females 382 (60.1%)			Males 254 (39.9%)		
Gender sample sizes	Females 44 (58.7%)			Males 31 (42.3%)			Females 172 (60.9%)			Males 110 (39.1%)		
Age sample sizes	a 1	b 6	c 37	a 1	b 2	c 28	a 7	b 26	c 139	a 11	b 2	c 97

Key: a=Age ≤ 10 years (Children); b=10 < Age ≤19 (Adolescents); and c=Age > 19 years (Adults)

An optimum sample size was considered since taking too large sample size implies a waste of resources while too small sample reduces the usefulness of the results (Shebeshi, 2011). Table 1.6.1 shows that the study is focused on a sample of 357 HIV+ patients from the two hospitals; with ages ranging from 0 to 67 years and averaging 32 years old. The descriptive statistics on gender and age distributions from the table are: 216 (60.5 %) females, 20 (5.6%) children, 36 (10.1%) adolescents and 301 (83.3%) adults. Furthermore, 75 (21%) of the HIV+ terminal patients were from Carolina hospital while 282 (79%) were from Embhuleni hospital.

Table 1.6.2 shows the distribution of died patients throughout the follow-up time of 7.5 years. The greatest number of patients (18/42 = 42.9%), died within the interval from 6 months to 1 year after ART initiation. Most patients (61.9%) died within 1 year from ART initiation. As

from the end of the 2nd year to the 7.5th year after ART initiation, the number of deaths was insignificant in comparison to the deaths which occurred up to the end of 2nd year. The mean and median survival times in this study were 1000.98 days/33 months/2.8 years and 548 days/18 months/1.5 years, respectively.

Table 1.6.2: Distribution of dead patients over Follow-up time (in years)

Time (years)	**≤0.5**	**0.5-1**	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-7.5**
No. dead	8	18	11	1	0	1	0	2	0
% dead	19.0	42.9	26.2	2.4	0	2.4	0	4.8	0
Cumulative No. died	8	26	37	38	38	39	39	41	41
%Cumulative No. died	19.0	61.9	88.1	90.5	90.5	93.0	92.9	100	100

Key: 2-3 means 2 < Time (in years) ≤ 3

Table 1.6.3 shows the distribution of dead and censored patients based on gender, hospital and on age groups. Table 1.6.3 shows that: out of the 41 patients who died, 20 (48.8%) were females; 21 (51.2%) males; and 1 (2.5%) children; 3 (7.3%) adolescents; and 37 (90.2%) were adults.

Table 1.6.3: Distribution of dead and censored patients by Gender, Age group and Hospital in the sample

	Gender		**Age group**			**Hospital**	
Status	**Females**	**Males**	**Children**	**Adolescents**	**Adults**	**Carolina**	**Embhuleni**
No. dead	20 (48.8%)	21 (51.2%)	1 (2.4 %)	3 (7.3%)	37 (90.3)	4 (9.8%)	37 (90.2%)
Censored	196 (62%)	120 (38%)	19 (6%)	33 (10.4%)	264 (83.6%)	71 (22.5%)	245 (77.5%
Total	216	141	20	36	301	75	282

Furthermore, 4 (9.8%) deaths occurred at Carolina hospital while 37 (90.2%) of deaths occurred at Embhuleni hospital. In addition to what is shown in Table 1.6.3, Table 1.6.4 shows the distribution of died patients based on other categorical variables. Table 1.6.4 shows the distribution of the 41 patients who died according to the respective categorical factor levels.

Table 1.6.4: Distribution of dead patients by ART adherence, TB history, WHO Stage and Treatment (Regimen 1) in the sample

Factor	**ART Adherence**			**TB history**		**WHO Stage**				**Treatment (Regimen 1)**				
Levels	poor	fair	good	yes	no	I	II	III	IV	1	5	6	18	19
No. dead	20	19	2	15	26	0	2	33	6	11	27	2	1	0
% dead	48.8	46.3	4.9	36.6	63.4	0	4.9	80.5	14.6	26.8	65.9	4.9	2.4	0

Key: Treatment (Regimen 1) levels: 1= NVP+D4T+3TC; 5= EFV+D4T+3TC; 6= EFV+AZT+3TC; 18= EFV+3TC+TDF, 19= NVP+3TC+TDF

1.6.5 Variables in the data set

The predictor variables (independent variables) presumed to be associated with HIV/AIDS are listed in Table 1.6.5. The independent variables are grouped into categorical and continuous covariates. The categorical variables in this study are in turn divided into nominal variables (for example, marital status) and ordinal variables (for example, WHO stage). The response or dependent variable in this study is the length of the survival time of an HIV+ terminal patient. The definitions of the highly technical predictor variables are given in Appendices H and I.

Table 1.6.5: List of variables in the data set

ID	Variable Name	Description	Codes/ Values /Description
1	Gender	Gender	female=1, male=2
2	Hospital	Name of hospital	Carolina=1, Embhuleni=2
3	HIVdisclosure	HIV disclosure status	disclosed=1, not disclosed=2
4	WHOS	WHO stage	I=1, II=2, III=3, IV=4
5	Maritalstatus	Marital status	single=1, married=2, staying together=3, widowed/separated/divorced=4
6	Treat1 (ART regimen one)	Antiretroviral drugs for regimen one	1 = NVP+D4T+3TC, 5 = EFV+D4T+3TC, 6 = EFV+AZT+3TC, 18 = EFV+3TC+TDF, 19 = NVP+3TC+TDF
7	ARTadherence	ART adherence	poor=1, fair=2, good=3
8	TBhistory	TB history	yes=1, no=2
9	ARTstart	Date ART started	Record of the date patient started ART as mm/dd/yy
10	Age	Age	Age of the patient on 01/01/2010
11	BaselineMass	Baseline mass	Mass of the patient on ART initiation date
12	FollowupMass	Follow-up mass	Average mass of the patient during follow-up period
13	BaselineCD4	Baseline CD4 cell	CD4 count on ART initiation
14	FollowupCD4	Follow-up CD4 cell	Average CD4 count of the patient during follow-up
15	BaselineHaemoglobin	Baseline haemoglobin	Haemoglobin level at ART initiation
16	FollowupHaemoglobin	Follow-up haemoglobin	Average haemoglobin level during the follow-up period
17	BaselineLymphocyte	Baseline lymphocyte	Lymphocyte level at ART initiation
18	FollowupLymphocyte	Follow-up lymphocyte	Average lymphocyte level during the follow-up period
19	BaselineWBC	Baseline White Blood Cell count	White blood cell count at ART initiation
20	FollowupWBC	Follow-up White Blood Cell count	Average white blood cell count during the follow-up period
21	BaselineviralLoad	Baseline viral load	Viral load at ART initiation
22	FollowupviralLoad	Follow-up viral load	Average viral load during the follow-up period
23	BaselineCreatinine	Baseline creatinine	Creatinine level at ART initiation
24	FollowupCreatinine	Follow-up creatinine	Average creatinine level during the follow-up
25	BaselineTotProtein	Baseline total protein	Total protein level at ART initiation
26	FollowupTotProtein	Follow-up total protein	Average protein level during the follow-up period
27	BaselineSodium	Baseline sodium	Sodium level at ART initiation
28	FollowupSodium	Follow-up sodium	Average sodium level during the follow-up period
29	BaselineALT	Baseline ALT	Alanine transaminase level at ART initiation
30	FollowupALT	Follow-up ALT	Average alanine transaminase during the follow-up
31	TimetoKidneyDamage	Time to kidney damage	Time it takes a patient to develop kidney damage after ART.
32	LENFOLdays / Survival time	Length of follow-up/ survival time in days	Length of follow-up time/survival time in days from ART initiation until transfer, lost to follow-up, death or study termination.

1.7 Limitations and delimitations of the study

The study has the following limitations.

- Results may not be generalizable beyond the population of Albert Luthuli municipality.
- Due to the missing data from some patient files and from other hospital registers, results may be biased if the missing mechanism of the data is not random (Soley-Bori, 2013).
- The sample size for the study was not adjusted for random loss to follow-up, random entry, hazard rate, length of accrual period, length of follow-up period or the number of covariates. Sample size determination formula for survival analysis (Tai et al., 2018; Kleinbaum and Klein, 2012; Chow et al., 2008; Cochran, 1977) and a sample size determination calculator software or statistical software packages for survival analysis (HyLown Consulting LLC, 2020; Ronan, 2018) could have yielded presumably a bigger sample.

The study has the following delimitation:

- The study is limited to Albert Luthuli municipality because of financial and time constraints which restricted the sample size and consequently the results of the research may not be generalizable to Gert Sibande district.

1.8 Assumptions

The study assumed the following.

- The patient files and ART electronic SOZO database were professionally compiled with minimal errors, if any.
- The research tool measures what it purports to measure when applied to the data for this study.
- HIV/AIDS trends at Carolina and Embhuleni Hospitals are a true representation of the HIV/AIDS trends in Albert Luthuli Municipality.
- Hospital blood tests, recordings and readings in Hospital were done with minimal errors, if any.
- The future can be predicted from the past.

1.9 Ethical considerations

The ethical approval for this study was granted by UNISA Ethics Review Committee (ERC) and (2017/SSR ERC/005) was given as reference number as evidenced in Appendix O. The permission to conduct the study at Carolina and Embhuleni hospitals was obtained from Mpumalanga Department of Health and (MP_201708_013) was given as the reference number as evidenced in Appendix N. The names of the patients were anonymously treated, and other data related to the patient was handled with utmost confidentiality in all the stages of the research. In addition, no reference to an individual respondent was recorded and all results are published in aggregate format. The electronic documents carrying confidential information on patients are all protected by some encryption and will be destroyed as per research policy.

1.10 Summary

The trends in HIV/AIDS and ART administration globally, in South Africa and in Albert Luthuli were covered in this chapter. The previous studies on survival analysis of HIV/AIDS patients, inside and outside Africa constitute this chapter. This chapter described the following items: problem statement, research question, purpose of the study, significance of the study and study assumptions. Some methodology items in this chapter are study area and hospitals, study design, data sources and data cleaning, sampling procedure and demographics and variables in the data. Ethical considerations and study limitations and delimitations are as well included in this chapter

1.11 Organisation of the study

Chapter two is devoted to the reviewing of literature on non-parametric analysis methods, survival analysis techniques, Cox regression, Logistic regression, Quantile regression and it also handles literature review on diagnostics of outliers and influential observations. Chapter 3 presents results analysis for the study. Chapter four presents study findings and discussions based on the results in chapter 3. Chapter five is the last chapter, and it summarises the study findings, presents recommendations, and gives suggestions for further research.

CHAPTER 2

REVIEW OF SURVIVAL ANALYSIS METHODS

This chapter reviews literature on how to use survival analysis methods in identifying factors that affect the survival of HIV+ terminal patients. Firstly, definitions of survival time, the observation window and censoring, survivor and hazard functions are presented. This is followed by reviews of:

- nonparametric methods (Kaplan-Meier estimator and the log-rank test for comparing of survivor functions);
- semiparametric methods (Cox proportional hazards regression modelling); and
- complementary regression methods (quantile and logistic regression modelling).

2.1 Introduction

In this study, survival time is defined as the time from ART initiation until an HIV+ terminal patient experiences death or until censoring or conclusion of the study without experience of an event. The analysis of the group data on survival time is called survival analysis (Goel et al., 2010). As described in Goel et al. (2010), Etikan et al. (2017) and, Etikan and Babatope (2018), the three assumptions used in survival analysis are; firstly, the participants who drop out or are censored have the same survival prospects as those who continue to be followed. Secondly, it is assumed that the survival probabilities are the same for the participants recruited early and late in the study. Thirdly, it is assumed that the event occurs at the time specified.

Two basic concepts which must be taken into cognizance before in-depth coverage of this work are observation window and censoring. The measurement window or observation window is the time period during which the researcher makes their observation (Perrigot et al., 2004). In this study the observation window is 7.5 years which was determined mainly by the availability of data. According to Perrigot et al. (2004), the length of the observation period is a personal and arbitrary judgement by the researcher(s). That is, there is little theoretical or empirical evidence to use as a guide for the choice of the length of the observation period. This is even though the interpretation of study results can vary with the length of the observation period of the study. According to Riddlesworth (2011), the term censoring was first used by Hald (1949) to describe observations in a study which contain incomplete information. As cited in

Windsperger et al. (2012), Li (1995-1996) defined censorship as an incomplete survival time like the lack of start date (left-censored), the missing date of an event (right-censored) or the loss from or disappearance from the sample within the measurement window (right-censored). Right censoring occurs when the study ends before a patient experiences an event or when a patient leaves the study before an event occurs. This study is characterised by right censoring where '1 = event of interest' (death of a patient due to HIV/AIDS related cause), while '0 = censored patient' (patient is alive at the end of study). The patient who is alive at the end of the study is assumed to have been attached to one hospital throughout the study or to have been lost to follow-up or to have been transferred to another hospital.

In survival analysis, estimation of the survivor and the hazard functions is of interest. The survivor function reflects the cumulative survival probabilities throughout the observation period (from ART initiation to death or censoring in this study), while the hazard function provides the rate of event occurrence (risk of death as regards this study) at a specific time (Perrigot et al., 2004; Etikan & Babatope, 2018). In the context of this study, suppose that T is a random variable representing the survival time of an HIV+ terminal patient during the observation period. Furthermore, suppose that $F(t)$ and $f(t)$ are the distribution and probability density functions of T, respectively. Then, the respective survivor and hazard functions of interest are (Box-Steffensmeir et al., 2004):

$$S(t) = P(T \geq t) = 1 - F(t) \; and \; h(t) = \frac{f(t)}{S(t)} = -\frac{d[lnS(t)]}{dt}. \qquad (1)$$

Furthermore, the respective mean and the median survival times are given by:

$$E(T) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt \; \; and \; S^{-1}(0.5). \qquad (2)$$

Also, of interest, can be the cumulative hazard function given by (Klein & Moeschberger, 2003; Kleinbaum & Klein, 2012):

$$H(t) = \int_0^t h(u)du = -ln\,[S(t)]. \quad (3)$$

2.2 Nonparametric survival analysis methods

2.2.1 The Kaplan-Meier estimator of the survivor function

In the context of this study, let: $0 < t_1 < t_2 < t_3 < \ldots < t_k < \infty$ be k observed times of death of patients in the cohort of HIV+ terminal patient during the observation period; $d_1, d_2, d_3, \ldots, d_k$ be the respective number of deaths at each of these times; and $n_1, n_2, n_3, \ldots, n_k$ be the corresponding number of remaining patients in the cohort at the respective times. The Kaplan-Meier (KM) estimator or the Product-Limit estimator of the survivor function $S(t)$ when death times are tied is given by (Etikan et al., 2017; Cleves et al., 2010):

$$\hat{S}(t) = \prod_{i|t_i \le t} 1 - \left(\frac{d_i}{n_i}\right). \quad (4)$$

The variance of the KM estimator is given by (Klein & Moeschberger, 2003):

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}. \quad (5)$$

However, confidence intervals of $S(t)$ are based on the asymptotic distribution of $\ln\{-\ln \hat{S}(t)\}$ whose estimate of the asymptotic variance is given by (Kalbfleisch & Prentice, 2002; Klein & Moeschberger, 2003):

$$\hat{\sigma}^2_{KM}(t) = \frac{\sum_{i|t_i \le t} \frac{d_i}{n_i(n_i - d_i)}}{\left\{\sum_{i|t_i \le t} ln\left(\frac{n_i - d_i}{d_i}\right)\right\}^2}. \quad (6)$$

Hence, the confidence bounds of $S(t)$ are calculated as:

$$\{\hat{S}(t)\}^{exp\left\{\pm z_{\frac{\alpha}{2}}\hat{\sigma}_{KM}(t)\right\}}, \quad (7)$$

where $z_{\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$ quantile of the standard normal distribution.

The Nelson-Aalen estimator of the cumulative hazard and the estimator of its variance are (Klein & Moeschberger, 2003):

$$\tilde{H}(t) = \begin{cases} 0, & if\ t \leq t_1 \\ \sum_{t_i \leq t} \frac{d_i}{n_i}, if\ t_1 \leq t' \end{cases} \quad (8)$$

and

$$\hat{\sigma}_H^2(t) = \sum_{t_i \leq t} \frac{d_i}{n_i^2}. \quad (9)$$

The survivor function $S(t)$ can also be estimated from the Nelson-Aalen estimate of the cumulative hazard $H(t)$ by substituting model (8) in model (3).

Remark: As reported in Etikan et al. (2017), for each given time interval, the survival probability is calculated as the number of surviving patients divided by the number of patients still at risk. Patients who died are not included in the denominator.

2.2.2 Comparison of survivor functions

The cohort of HIV+ terminal patients in this study is not homogeneous with respect to their characteristics that may affect their survival. Hence, it will be necessary to test the equality of survivor functions among groups of patients. That is, to test the null hypothesis of the form

$$H_0: S_1(t) = S_2(t) = \cdots = S_r(t) \equiv h_1(t) = h_2(t) = \cdots = h_r(t), \quad (10)$$

for $t \geq 0$, where $S_i(t)$ and $h_i(t)$ are the respective survivor and hazard functions of the i^{th} group patients, and r is the number of groups. The alternative hypothesis of most interest is that the survival time of one group is stochastically bigger or smaller than the survival time for the other group, and is given by:

$$H_1: either\ S_i(t) \geq S_j(t), or\ S_i(t) \leq S_j(t)\ for\ some\ i \neq j. \quad (11)$$

A vertical gap between two graphs of the survival functions shows that at a given period of time, one group had a higher probability of surviving while a horizontal gap shows that one group took longer to experience an event (Etikan et al., 2017).

The log-rank test is the most commonly used statistical test for comparing survival functions of two or more groups (Clark et al., 2003; Etikan et al., 2017). A very important assumption for the appropriate use of the log rank test (and the Cox PH regression model) is that the hazards are proportional over time and this implies that the effect of a risk factor is constant over time (Sullivan, 2016).

Comparison of survivorship experience between two or more groups in this study is done by graphing the Kaplan-Meier estimator of the survivorship functions. The statistical difference between the survival curves is then tested using Log-rank test, generalised Wilcoxon test, Taron-Ware test, Peto-Prentice test or Harrington-Fleming test. According to Zhang (2000), Log-rank test is the most powerful test which puts equal weight to all times, while Wilcoxon test places more weight to early survival times than to later times.

In the context of this study and times in (section 2.2.1), let d_{ij} and n_{ij} $(i = 1,2, \dots, r\ ;\ j = 1,2, \dots, k)$ be the respective number of deaths and remaining number of patients in the i^{th} group at the j^{th} time in the cohort of patients. Furthermore, let $d_j = \sum_{i=1}^{r} d_{ij}$ and $n_j = \sum_{i=1}^{r} n_{ij}$. Then, the expected number of deaths in the i^{th} group at time t_j is $e_{ij} = \frac{n_{ij} d_j}{n_j}$ (Hosmer et al., 2008).

Hence, the chi-square test statistic, with a chi-square distribution with $r-1$ degrees of freedom, for comparing the group survivor functions is given by (Cleves et al., 2010):

$$\chi = \boldsymbol{u}^T \boldsymbol{V^{-1}} \boldsymbol{u} \sim \chi^2_{r-1}, \tag{12}$$

where the r row vector

$$\boldsymbol{u}^T = \sum_{j=1}^{k} w(t_j)\left(d_{1j} - e_{1j}, d_{2j} - e_{2j}, \dots, d_{rj} - e_{rj}\right), \tag{13}$$

and the *(il)*th element of the $r \times r$ covariance matrix V is given by:

$$v_{il} = \sum_{j=1}^{k} \frac{w^2(t_j) n_{ij} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{il} - \frac{n_{ij}}{n_j}\right), \quad i = 1,2,\dots,r; l = 1,2,\dots,r; \ \delta_{il} = 1, \tag{14}$$

if $i = l$ and 0 otherwise. The $w(t_j)$ is a positive function equal to zero when $n_{ij} = 0$, and obtains the log-rank test statistic if equal to 1 when $n_{ij} > 0$ (Klein et al., 2003; Cleves et al., 2010). H_0 is rejected if $\chi > \chi^2_{r-1,\alpha}$ or if the *p-value* < α (*r* is number of groups and α is the level of significance of the test).

Remark: Other chi-square test statistics are obtained by choice of $w(t_j)$. For example, $w(t_j) = n_j$ obtains the Wilcoxon chi-square test statistic (Karadeniz & Ercan, 2017).

The Log-rank test is primarily a significance test and it does not estimate the magnitude of the difference between the survival experiences (Johnson & Shih, 2012). In order to assess multiple variables, to adjust for both categorical and continuous prognostic factors and to estimate the magnitude of the difference between the survival experiences, a more complex method such as the Cox regression model is needed (Johnson & Shih, 2012). Unlike the Log-Rank test, Cox regression estimates the magnitude of the difference between the survival experiences through the hazard ratios.

2.3 Semiparametric survival analysis methods

2.3.1 The Cox proportional hazards (PH) model

2.3.1.1 The theoretical aspects of the Cox PH model

The objective of the study is to identify factors that affect the survival of HIV+ terminal patients. The Cox proportional hazards (PH) model expresses the patient hazard rates as functions of potential such factors (covariates) as follows. Let $\boldsymbol{X}_i^T = (X_{i1}, X_{i2}, \dots, X_{ip})$ be a p-dimensional vector of the values of the covariates associated with the i^{th} patient. Then, the Cox proportional hazards regression model is as follows (Klein & Moeschberger, 2003; Etikan & Babatope, 2018):

$$h_i(t) = h_0(t)exp\{\boldsymbol{X}_i^T\boldsymbol{\beta} = \sum_{j=1}^{p}\beta_j X_{ij}\}, \quad (15)$$

where $\boldsymbol{\beta}^T$=$(\beta_1, \beta_2, \dots, \beta_p)$ is a p-dimensional vector of regression coefficients to be estimated from the data, and $h_0(t)$ is the unspecified baseline hazard function that does not have to be estimated. The hazard model in model (15) makes no assumptions about the shape of the hazard function over time. A hazard function could be constant, increasing, decreasing, or it could be a combination of two or three of these graph trends.

Model (15) can be written in terms of the survivor function (Kleinbaum & Klein, 2012) :

$$S_i(t) = S_0(t)^{exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}. \quad (16)$$

The model (15) assumptions which may be violated by the data, are: (i) the covariates $\boldsymbol{X}_i^T$ do not vary with time and hence the hazard rate ratios of pairs of patients do not vary with time; (ii) censoring and survival are independent; and (ii) the log hazard rate is indeed a linear function of the covariates.

Non-parametric methods (for example Kaplan-Meier) and semi-parametric (for example Cox proportional hazards model) are not the only approaches used to analyse censored time to event data or survival time data. The parametric models form a good alternative to Cox PH model

when the PH assumptions fail. One way of writing parametric models is as the Cox PH model given by equation (15) but with the baseline hazard function determined by the assumed distribution of the survival time (Hamidi et al., 2017). For example, the baseline hazard function is constant over time if the assumed survival time distribution is exponential. Other widely used survival data distributions are Weibull, gamma, log-normal, log-logistic and normal. Nakhaee and Law (2011) used four parametric survival models (exponential, Weibull, log-normal and log-logistic) for survival analysis of HIV/AIDS data in Australia, and the Weibull model was found to be the best parametric model. Parametric models are known to be more efficient than non-parametric models when using survival analysis to make projections about the risk of death (May et al., 2003), and future trends in mortality (Veugelers et al., 1998). Although the Cox model is frequently used in survival analysis, parametric models may fit data better and give more precise estimates of the quantities of interest (Hamidi et al., 2017). Solomon and Hutton have suggested that accelerated life models may be useful since there is some robustness to misspecification of survival regression models (Kwong & Hutton, 2003).

2.3.1.2 The application of the Cox model

The Cox PH model allows for prognostic factors. In the context of this study and in medical literature, a prognostic factor is a variable that can be used to estimate the chance of recovery from a disease. Richardson (2009) defines a prognostic model as a predictive tool whose purpose is to predict the level of increase in a beneficial effect, or the decrease in risk of some adverse event, for instance, death. This definition concurs with Nakhaee and Law (2011), who view predictions on progression of a disease as helping clinicians to estimate trends in their patient outcomes. This means a prognostic model can assist a health practitioner in making clinical decisions, for example in trying to determine which patients might benefit from a treatment or therapy in view of some clinical and demographic factors. The decision on the treatment to be given to a patient may be based upon evidence obtained through the application of a prognostic model. Therefore, it is important to produce a reliable and an accurate model (Richardson, 2009).

2.3.2 Inferences about the Cox model parameters

Consider model (15) in the previous section. The parameters are estimated as values of β which maximize the Cox likelihood (also called partial likelihood) function for censored data. The partial likelihood function is as follows (Hosmer et al., 2008; Cleves et al., 2010):

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \left(\frac{exp\left(\boldsymbol{X}_j^T \boldsymbol{\beta}\right)}{\sum_{i \in R_j} exp(\boldsymbol{X}_i^T \boldsymbol{\beta})} \right). \quad (17)$$

where R_j is the group of patients at the risk of death at time t_j. In particular, the maximum likelihood estimate of β, which is $\hat{\beta}$, is found by iteratively solving the equations $\frac{\partial \ln L(\beta)}{\partial \beta}|_{\beta=\hat{\beta}} = 0$. The popular iteration algorithms are the Fisher's scoring (Storvik, 2011) and Newton-Raphson algorithms (Zhou, 2016). The estimate of the covariance of $\widehat{\boldsymbol{\beta}}$ is a function of the inverse of the matrix $\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$, and is of the form:

$$\widehat{\boldsymbol{Cov}}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}^T \widehat{\boldsymbol{W}}\, \boldsymbol{X})^{-1}, \quad (18)$$

where $\widehat{\boldsymbol{W}}$ is a diagonal weight matrix, and $\boldsymbol{X}$ is the design/incidence matrix whose i^{th} row is $\boldsymbol{X}_i^T$. The standard error of $\hat{\beta}_j$ is the square root of the j^{th} diagonal element of model (18).

Inferences about $\boldsymbol{\beta}$ can be made using the partial likelihood ratio test and/or the Wald test. Firstly, consider testing the overall goodness of fit of a fitted Cox PH model with *p*-vector parameter $\boldsymbol{\beta}$. The fitted model contains a size *p* subset of all possible explanatory variables and hence is called a reduced model as opposed to a full model (or saturated model) which contains all possible explanatory variables. Then the hypotheses about the goodness of fit of the Cox PH model are given by:

H_0: *the reduced model fit is good* versus H_1: *the reduced model fit is notgood*.

The partial likelihood ratio test statistic for the hypotheses compares the log partial likelihood ratios of the reduced and full models as follows. Consider the log partial likelihood function $l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})$. Let $l\left(\widehat{\boldsymbol{\beta}}\right)$ be $l(\boldsymbol{\beta})$ evaluated at the maximum likelihood estimate of $\boldsymbol{\beta}$ for the reduced model, and let $l(\boldsymbol{t})$ be $l(\boldsymbol{\beta})$ but evaluated at the maximum likelihood estimate of $\boldsymbol{\beta}$ of the full/saturated model. Then the partial likelihood ratio test statistic is (Zhang, 2015):

$D = 2\{l(\boldsymbol{t}) - l(\widehat{\boldsymbol{\beta}}) \sim \chi^2_{n-p}$ (asymptotically). (19)

H_0 is rejected if $D > \chi^2_{n-p,\alpha}$ or if the *p-value*<α (*n* is number of patients; *p* is number of parameters and α is the level of significance of the test). The Wald test statistic for testing the hypotheses H_0: $\boldsymbol{\beta} = 0$ versus H_1: $\boldsymbol{\beta} \neq 0$ is (Hurlin, 2015):

$W = \widehat{\boldsymbol{\beta}}^T [\widehat{\boldsymbol{Cov}}(\widehat{\boldsymbol{\beta}})]^{-1} \widehat{\boldsymbol{\beta}} \sim \chi^2_{n-p}$ (asymptotically). (20)

H_0 is rejected if $W > \chi^2_{n-p,\alpha}$ or if the *p-value*<α. To test the null hypothesis H_0: $\beta_j = \beta_j^0$, the Wald test statistic $Z = \frac{\hat{\beta}_j - \beta_j^0}{\text{se}(\hat{\beta}_j)}$ is used, where se $(\hat{\beta}_j)$ is the estimate of the asymptotic standard error of $\hat{\beta}_j$ (square root of the j^{th} diagonal element of $\widehat{\boldsymbol{Cov}}(\widehat{\boldsymbol{\beta}})$). Under the null hypothesis, the asymptotic distribution of *Z* is the standard normal distribution.

2.3.3 Model selection and diagnostics in Cox PH models

2.3.3.1 Model and variable selection

If the assumptions of model (15) are not violated by the data, then developing a Cox PH model involves variable selection. To do this, the partial likelihood ratio and the Wald tests are used in conjunction with the information criteria such as Akaike's (1973) Akaike information criterion (AIC) (Xu et al., 2009):

$AIC = -2\, ln\, L(\widehat{\boldsymbol{\beta}}) + 2p,$ (21)

where *p* is the number of model parameters. The model with the smallest AIC among competing Cox PH models is the best.

Regarding model selection, Hosmer et al. (2008) suggested the consideration of issues such as clinical importance and adjustment for confounding, as well as statistical significance. In the context of this study, the variable selection process outlined by Hosmer et al. (2008) follows.

Step 1: Fit a Cox PH model containing all possible variables and use the Wald tests to determine the variables with significant effects, at the 5% level of significance, to be retained in the model. Retain in the model other variables not selected with this criterion but which appear to be of clinical importance.

Step 2: Refer to the partial likelihood ratio test to compare the original Cox PH model (Full model) and Cox PH model after deleting insignificant variables at step 1 (Reduced model). That is to confirm that the set of deleted variables are jointly insignificant but may be important.

Step 3: After fitting the reduced model, do an assessment on whether removal of the covariate has produced an "important" change in the coefficients of the variables remaining in the model. In general, the value of about 20 percent is used as an indicator of an important change in a coefficient. If the variable excluded is an important confounder, it should be added back into the model. This process is continued until no covariates could be deleted from the model.

Step 4: All variables excluded from the initial multivariable model are added, one at a time, to ascertain that they are neither statistically significant nor important confounders.

Step 5: Testing the linearity of the continuous covariates. The objective of linearity test is to determine whether the data support the hypothesis that the effect of the covariate is linear in the log hazard.

Step 6: The need for interactions in a model is determined in this step. The formation of some possible interaction terms from the main effects in the model is done. Clinically important and statistically significant interactions are considered. Each of the individual interaction created is assessed by comparing the model carrying the interaction term to the main effects model by referring to the partial likelihood ratio test. All interactions which are found to be significant at the 10% level are added jointly to the main effects model. Wald statistic p-values are used as a guide to eliminate interactions from the model, while significance is checked by the partial likelihood ratio test. The model at the conclusion of this step is referred to as the *preliminary model.*

Step 7: This step produces the *final model* which is thoroughly evaluated through the checking for the adherence to key model assumptions. Diagnostic statistics is done to check for influential observations and testing for overall goodness of fit. According to Harrell Jr. (2018), the following are identified as potential problems in the modelling process: violation of assumptions, omission of important predictors, missing data, incorrect imputation and over-

fitting. In view of the problems in the modelling process stated above, this study intends to make several simplifications and assumptions in fitting the statistical model. The real-world situations are often too complex to model without such assumptions and simplifications.

2.3.3.2 Proportional hazards and linearity of the predictor assumptions

The null hypothesis for the linearity test is that the predictor in the Cox PH model is $\boldsymbol{X}_i^T\boldsymbol{\beta}$. The hypothesis may be tested by testing the null hypothesis that $\theta_2 = 0$ in the Cox PH model (Cleves et al., 2010):

$$h_i(t) = h_0(t)exp\{\theta_1(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}) + \theta_2(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})^2\}, \tag{22}$$

where $\widehat{\boldsymbol{\beta}}$ is from fitting Cox PH model (15). Rejecting the null hypothesis implies that $\boldsymbol{X}_i^T\boldsymbol{\beta}$ is an incorrect specification of the predictor in the Cox PH model. In addition to Wald test, the family of quadratic form tests which include the weighted Log-rank test, proportionality test (Lin et al., 2006), and the test proposed by Seagusa et al. (2014) can be used to test the null hypothesis that $\theta_2 = 0$.

The proportional hazards assumption may be tested by testing the null hypothesis that $\theta_2 = 0$ in the Cox PH model (Cleves et al., 2010):

$$h_i(t) = h_0(t)exp\{\theta_1(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}) + \theta_2(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})t\}, \tag{23}$$

where $\widehat{\boldsymbol{\beta}}$ is from fitting Cox PH model (15). Rejecting the null hypothesis implies that the proportional hazards assumption does not hold. When the PH assumption holds and the true model is the Cox model then Wald test and log-rank test tend to be the most powerful tests while other testing procedures such as the two-stage procedure and Gray's (1994) score tests, are often powerful for certain alternatives (Seagusa et al.,2014).

The 'ASSESS' optional statement of PROC PHREG in SAS Version 9.4 is useful for checking the functional forms of covariates, and for checking the proportional hazards assumption (Lin

et al., 2002; University of California, Los Angeles, Statistical Consulting Group, 2019). The ASSESS statement plots the cumulative score residuals against the values of the covariate to check functional form, and against time to check proportional hazards (PH) assumption. The RESAMPLE optional statement, also in PROC PHREG, computes the p-value of the Kolmogorov-type supremum test for model goodness of fit using a sample of 1,000 simulated residual patterns. The process then generates supposedly identical graphical and numerical results for both the functional form and the PH model. The null distribution of the cumulative martingale residuals can be simulated through zero-mean Gaussian processes. If the observed pattern (path from the actual data) is within the simulated cloud of random paths or the supremum test (numerical statistical test) is non-significant ($p - value > 0.05$), it is concluded that the functional form holds; otherwise the functional form is violated. As for proportional hazards assumption; if the observed pattern (path from the actual data) is within the simulated cloud of random paths or the supremum test (numerical statistical test) is non-significant ($p - value > 0.05$), it is concluded that that the proportional hazards assumption holds; otherwise the assumption is violated.

Alternatively, a graph of the scaled Schoenfeld residuals versus time can be used to check the proportional hazards assumption. The Schoenfeld residual for a continuous covariate X_{ij} ($j = 1,2, \dots, p$) and for the observation i from which the scaled Schoenfeld residual (r_{ij}) is obtained (Cleves et al., 2010):

$$r_{ij} = X_{ij} - \frac{\sum_{i \in R_j} X_{ij} \exp\{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}\}}{\sum_{i \in R_j} \exp\{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}\}}, \quad (24)$$

where R_j is the group of patients at the risk of death at time t_j. The graph of r_{ij} versus t_j has approximately zero slope if the proportional hazards assumption holds.

2.3.3.3 Outliers and influential observations

An outlier is defined as an observation that has large standard or studentized residuals (high difference between observed value and predicted value) (Chatterjee & Hadi, 1986). As outlined by Viljamaa (2017), outlier detection is used to detect and, when needed, to remove anomalous observations from data. According to Viljamaa (2017) and in terms of Tukey's

(1977) fences as limit, the simplest single-step outlier detection method for univariate data is the boxplot. Using the boxplot method, outliers are observations below $Q_1 - 1.5(IQR)$ or above $Q_3 + 1.5(IQR)$, where the interquartile range $(IQR) = Q_3 - Q_1$, and Q_k is the k^{th} quartile of the data. Outliers can greatly affect the inferences from the data and therefore removing them from analysis improves the inferences. According to Budzier and Flyvbjerg (2013), rejection or exclusion of outliers is the most common method to deal with outliers, and this is common in most academic studies.

An influential observation is that, if removed, significantly changes estimates of the model's parameters. An extreme observation can be an outlier but not influential, it can be influential but not an outlier or it can be both an outlier and an influential observation. The effect of an influential observation on the regression model fit is measured in a variety of ways, both absolute and relative. A variety of residuals (score residuals, Schoenfeld residuals, delta-beta residuals) have been proposed as ways of quantifying and assessing influence (Cleves et al., 2010). Mack (2016) and Chan (2017) summarised the following as common ways (and their criteria of usage) of measuring the influence of the i^{th} observation in large samples (n>20, for Mack).

- Cook's distance (D): If Cook's distance, $D_i > \frac{4}{n}$, then the i^{th} observation is influential.
- Difference in beta (DFBETA): If the absolute difference in beta, $|\text{DFBETA}_\text{i}| > \sqrt{\frac{4}{\text{n}}}$, then the i^{th} observation is influential.
- Difference in fit (DFFITS): If the absolute difference in fits, $|\text{DFFITS}_\text{i}| > \sqrt{\frac{4\text{p}}{\text{n}}}$, then the i^{th} observation is influential (n is sample size while p is the number of parameters in the model).

According to Canette (2014), DFBETA is the amount that an estimate of a regression model of a particular parameter changes when an observation is suppressed. That is, if $\hat{\beta}$ is the estimate for parameter β obtained from the full data and $\hat{\beta}_{(i)}$ is the corresponding estimate obtained when the i^{th} observation is suppressed, then $DFBETA_i = \hat{\beta} - \hat{\beta}_{(i)}$. To find out if any of the observations has too much influence on the estimated parameters, graphs of the components of the $\hat{\beta} - \hat{\beta}_{(i)}$ versus observations i are used with the above DFBETA criterion.

The difference in fits for observation *i*, denoted $DFFITS_i$, is defined as (The Pennsylvania State University, 2018):

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_i h_{ii}}}, \quad (25)$$

where: the numerator is the difference between the predicted i^{th} responses from the estimated model using the full data with and without the i^{th} observation, respectively; MSE_i is the mean square error estimated without the i^{th} observation; and h_{ii} is the leverage for the i^{th} observation. The graphical analysis proceeds in the same way as in the DFBETA analysis only that DFFITS is a global measure of influence which is not tied to a particular covariate.

As pointed out by Cleves et al. (2010), there is no firm guidance regarding how to treat influential observations, but one thing that is certain is that observations should not be automatically removed from the model. According to Hosmer et al. (2008) and in congruence to this study, the final decision on the continued use of a subject's data to fit the model will depend on the observed percentage change in the coefficients that results from deleting the subject's data and, more importantly, the clinical plausibility of that subject's data. The inputs from diverse literature on the assessment of model adequacy point to the fact that the assessment process must be done repeatedly, exhaustively and meticulously.

2.3.3.4 Goodness of fit

A graph of the cumulative hazard of the Cox-Snell residuals versus the residuals can be used to check the goodness of fit of the model. The Cox-Snell residual for the i^{th} observation from which the cumulative hazard of the residuals is (Ansin, 2015; Cleves et al., 2010):

$$CSr_i = \hat{H}_0(t_i) exp\{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}\}, \quad (26)$$

where $\hat{H}_0(t_i)$ is an estimate of the cumulative baseline hazard rate which is obtained from the Cox PH model fit when all covariates are set to equal to zero ($\mathbf{X}_i^T = \mathbf{0}$ for all *i*). The graph of the cumulative hazard of the Cox-Snell residuals versus the Cox-Snell residuals is expected to approximately follow a 45^0 continuous straight line passing through $(0,0)$ with a slope

approximately equal to 1 if the conditional distribution of the cumulative hazard function given the covariate vector is reasonably well approximated by the exponential distribution.

To assess the predictive ability of the Cox PH model, Harrell's concordance C-index may be used. Harrell's C and the equivalent parameter Somers' D were proposed as measures of the general predictive power of a general regression model by Harrell et al. (1982) and Harrell et al. (1996) (Newson, 2010). According to Schmid et al. (2016), a value of Harrell's C = 0,5 corresponds to a non-informative prediction rule whereas Harrell's C = 1 corresponds to perfect association. According to Zhao (1998), there are at least three uses of measures of a model's predictive accuracy which are as follows. Firstly, to quantify the utility of a model, secondly to check the model for overfitting or lack of fit, and thirdly to rank competing models. The inputs from diverse literature on the assessment of model adequacy point to the fact that the assessment process must be done repeatedly, exhaustively and meticulously.

2.3.4 Interpretation of the coefficients of the Cox PH model

The hazard ratio (HR) associated with the j^{th} covariate is (Cleves et al., 2010):

$$HR_j = e^{\hat{\beta}_j}, j = 1,2, \dots, p, \tag{27}$$

where is $\hat{\beta}_j$ is the estimate of the coefficient of the j^{th} covariate (β_j). For continuous covariates, if $\hat{\beta}_j > 0$ then a unit increase in the j^{th} covariate increases the hazard by $(HR_j - 1)100\%$. Otherwise, a unit increase in the j^{th} covariate decreases the hazard by $(HR_j - 1)100\%$. For categorical variables, the j^{th} covariate is actually the j^{th} category/level of the categorical variable. Hence, if $\hat{\beta}_j > 0$ then patients in the j^{th} category/level face a hazard $(HR_j - 1)100\%$ greater than those in the specified reference category/level. Otherwise, the patients in the j^{th} category/level face a hazard $(HR_j - 1)100\%$ lower than those in the specified reference category/level. The $(1 - \alpha)100\%$ confidence interval for the true hazard ratio associated with the j^{th} covariate (e^{β_j}) is given by:

$$CR_j = e^{\hat{\beta}_j \pm z_{\frac{\alpha}{2}} se(\hat{\beta}_j)}, j = 1,2, \dots, p, \tag{28}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $se(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. If CR_j includes one, then there is no association between the hazard (and the survival) and the j^{th} covariate. In summary e^{β_j} indicates how large (or small) the hazard in one group or subject is with respect to the hazard in the reference group or subject. Summarised below is the interpretation of Cox PH model regression coefficients (McGready, 2009):

- β_j>0: Higher hazard (poorer survival) associated with j^{th} covariate since e^{β_j}>1,
- β_j<0: Lower hazard (better survival) associated with j^{th} covariate since e^{β_j}<1,
- β_j=0: No association between hazard (and survival) and j^{th} covariate, since e^{β_j}=1.

2.4 Logistic regression in survival analysis

Logistic regression analysis can be used to determine factors that significantly affect the odds of survival of HIV+ patients, and to predict the odds of survival of HIV+ patients from the given patient covariates. In the context of this study, for $i = 1,2,\dots,n,$ let the random variable $Y_i = 1$ if the i^{th} HIV+ patient dies within the observation period, and $Y_i = 0$ if otherwise. Thus, Y_i has Bernoulli distribution with mean $\pi_i = P(Y_i = 1)$ which is assumed to depend on the p-dimensional vector of the values of the covariates, $\boldsymbol{X}_i^T = (X_{i1}, X_{i2}, \dots, X_{ip})$, associated with the i^{th} HIV+ patient through some link function such as the probit, logit (equivalently log), etc. The most popular is the logit (log) link function (Dayton, 1992; Chatterjee & Chatterjee, 2010). Using the logit (log) link function, the logistic regression model is given by:

$$\pi_i = \frac{exp\left\{\boldsymbol{X}_i^T\boldsymbol{\beta}=\sum_{j=1}^{p}\beta_j X_{ij}\right\}}{1+exp\left\{\boldsymbol{X}_i^T\boldsymbol{\beta}=\sum_{j=1}^{p}\beta_j X_{ij}\right\}}, i = 1,2,\dots,n, \tag{29}$$

where $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \dots, \beta_p)$ is a p-dimensional vector of regression coefficients to be estimated from the data. The strong model assumptions are that the logit (log) link function is appropriate and that $\boldsymbol{\beta}$ does not vary with the patients. An equivalent expression of model (29) is (Dayton, 1992; El-Habil, 2012):

$$logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\beta}, i = 1,2,\dots,n. \tag{30}$$

The ratio

$$\frac{\pi_i}{1-\pi_i} = exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}, i = 1,2,\dots,n, \quad (31)$$

is called the odds of dying (Eckel, 2008). Thus, the other strong assumption of the logistic regression model that has been implicitly made above is the linearity of the predictor $\mathbf{X}_i^T\boldsymbol{\beta}$ (Tse, 1993).

2.4.1 Inferences about the logistic regression model parameters

The estimator of $\boldsymbol{\beta}$ is the maximum likelihood estimator is (Dayton, 1992):

$$\widehat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} \prod_{i=1}^{n} \left(\frac{exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}{1+exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}\right)^{Y_i} \left(\frac{1}{1+exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}\right)^{1-Y_i}. \quad (32)$$

The function being maximized is the likelihood function of $\boldsymbol{\beta}, L(\boldsymbol{\beta})$. Inferences about $\boldsymbol{\beta}$ are as were discussed for the Cox PH model (see section 2.3.2), and are repeated here for completeness. The estimate of the covariance of $\widehat{\boldsymbol{\beta}}$ is a function of the inverse of the matrix $\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$, and is of the form:

$$\widehat{\boldsymbol{Cov}}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}^T\widehat{\boldsymbol{W}}\,\boldsymbol{X})^{-1}, \quad (33)$$

where $\widehat{\boldsymbol{W}}$ is a diagonal weight matrix, and $\boldsymbol{X}$ is the design/incidence matrix whose i^{th} row is $\boldsymbol{X}_i^T$. The standard error of $\hat{\beta}_j$ is the square root of the j^{th} diagonal element of model (33). Inferences about $\boldsymbol{\beta}$ can be made using the likelihood ratio test and/or the Wald test (Dayton, 1992; Tse, 1993; Ijomah & Nwali, 2018). Firstly, consider testing the overall goodness of fit of a fitted logistic regression model with p-vector parameter $\boldsymbol{\beta}$. The fitted model contains a size p subset of all possible explanatory variables, hence it is called a reduced model, as opposed to a full model (or saturated model), which contains all possible explanatory variables. Then the hypotheses about the goodness of fit of the logistic regression model are given by:

H_0: *the reduced model fit is good* versus H_1: *the reduced model fit is notgood*.

The likelihood ratio test statistic for the hypotheses compares the log partial likelihood ratios of the reduced and full models as follows. Consider the log partial likelihood function $l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})$. Let $l(\widehat{\boldsymbol{\beta}})$ be $l(\boldsymbol{\beta})$ evaluated at the maximum likelihood estimate of $\boldsymbol{\beta}$ for the reduced model, and let $l(\boldsymbol{t})$ be $l(\boldsymbol{\beta})$ but evaluated at the maximum likelihood estimate of $\boldsymbol{\beta}$ of the full/saturated model. Then the likelihood ratio test statistic is given by:

$$D = 2\{l(\boldsymbol{t}) - l(\widehat{\boldsymbol{\beta}})\} \sim \chi^2_{n-p} \text{ (asymptotically).} \quad (34)$$

H_0 is rejected if $D > \chi^2_{n-p,\alpha}$ or if the *p-value* < α (*n* is number of patients; *p* is number of parameters and α is the level of significance of the test).

The Wald test statistic for testing the hypotheses H_0: $\boldsymbol{\beta} = 0$ versus H_1: $\boldsymbol{\beta} \neq 0$ is given by

$$W = \widehat{\boldsymbol{\beta}}^T [\widehat{\boldsymbol{Cov}}(\widehat{\boldsymbol{\beta}})]^{-1} \widehat{\boldsymbol{\beta}} \sim \chi^2_{n-p} \text{ (asymptotically).} \quad (35)$$

H_0 is rejected if $W > \chi^2_{n-p,\alpha}$ or if the *p-value* < α. To test the null hypothesis H_0: $\beta_j = \beta_j^0$, the Wald test statistic $Z = \frac{\hat{\beta}_j - \beta_j^0}{\text{se}(\hat{\beta}_j)}$ is used, where se $(\hat{\beta}_j)$ is the estimate of the asymptotic standard error of $\hat{\beta}_j$ (square root of the j^{th} diagonal element of $\widehat{\boldsymbol{Cov}}(\widehat{\boldsymbol{\beta}})$). Under the null hypothesis, the asymptotic distribution of *Z* is the standard normal distribution.

2.4.2 Model selection and diagnostics in logistic regression

2.4.2.1 Model and variable selection

As was discussed for the Cox PH model (see section 2.3.3), if the assumptions of equation (29) are not violated by the data, then developing a logistic model involves variable selection. To do this, the likelihood ratio and the Wald tests are used in conjunction with the information criteria such as the Akaike information criterion (AIC) as in model (21) and is (Ijomah & Nwali, 2018):

$$AIC = -2\,ln\,L(\widehat{\boldsymbol{\beta}}) + 2p, \tag{36}$$

where $L(\widehat{\boldsymbol{\beta}})$ is the maximum likelihood achievable by the model and p is the number of parameters of the model. The best model is the one which minimizes the AIC. Three commonly used methods for variable selection are forward selection, backward elimination, and stepwise selection (Tse, 1993). In this study, the stepwise purposeful variable selection procedure as was proposed by Hosmer et al. (2008) is used for variable selection. Stepwise purposeful variable selection procedure selects variables into the model as was discussed in (section 2.3.3.1) on Cox PH modelling.

2.4.2.2 Outliers and influential observations

The analysis of residuals and the identification of outliers and influential observations in Logistic regression are not studied so frequently to check the adequacy of the fitted model (Sarkar et al., 2011). However, the definitions on outliers and influential observations as in (section 2.3.3) apply to Logistic regression as well. The general rule of thumb is that the higher the value of Cook's D, the more influential the observation is. Additionally, if the Cook's D value is greater than $^4/_n$ (n is number of observations), the value is considered an outlier (Sarkar et al., 2011).

2.4.2.3 Appropriateness of the linearity of predictor

The logistic regression model (30) assumes a linear relationship between the independent variables and log odds (Statistics Solutions, 2019; Schreiber-Gregory & Jackson Foundation, 2018). If non-linearity is suspected, one solution is to create an interaction term of the independent variables times the natural log of that independent variable. If any of the terms is found to be significant (using say Wald test), then it suggests non-linearity in the logit.

2.4.2.4 Goodness of fit

A good model is one which 'fits' the data well, in the sense that the values predicted by the model are in close agreement with those observed (Giancristofaro & Salmaso, 2003). Among a wide range of measures of the goodness of fit of a logistic regression model is the Bayesian Information Criterion (BIC) given by (Liddle, 2008):

$$BIC = -2\,ln\,L(\widehat{\boldsymbol{\beta}}) + p\,ln\,n, \tag{37}$$

where n is the number of patients and p is the number of parameters in the model. The BIC is closely related to Akaike information criterion (AIC). The BIC judges a model by considering the distance between the fitted values and the corresponding true expected values, and the optimal model is the one that tends to have its fitted values closest to the true expected values (smallest BIC value) (El-Habil, 2012).

Receiver operating characteristics (ROC) curve and Hosmer-Lemeshow goodness of fit test are among the common ways of assessing the goodness of fit of the logistic regression model. The ROC graph is a plot of true positive rate (sensitivity) $\in (0,1)$ versus false positive rate (specificity) $\in (0,1)$ where the outcome probabilities are estimated/predicted using the fitted logistic model (Bewick et al., 2004). The ROC curve starts at coordinate (0,0) and ends at coordinate (1,1). The area under the ROC curve (AUROC), which varies from 0.5 (model with no predictive ability) to 1 (model with perfect predictive ability), is used to assess the predictive ability of a logistic regression model (Bewick et al., 2004, 2005). Thus, the larger the AUROC the higher is the predictive ability of a model.

The Hosmer-Lemeshow goodness of fit test is used to assess whether the number of expected events from the logistic regression model fit are close to the number of observed events in the data. The Hosmer-Lemeshow test statistic ($\widehat{HL}$) is (Guffey, 2012):

$$\widehat{HL} = \sum_{i=1}^{K} \frac{(O_i - n_i\bar{p}_i)^2}{n_i\bar{p}_i(1-\bar{p}_i)} \sim \chi^2_{K-2}, \tag{38}$$

where O_i is the number of observed events in group i, n_i is the number of observations in group i, $\bar{p}_i$ is the average predicted probability in group i and K is the number of groups. The logistic regression model poorly fits the data if $\widehat{HL} > \chi^2_{K-2,\alpha}$ or the *p-value* < α (the level of significance of the test).

The concordance of association between the response variable and the covariates and the Harrel's C-statistic are used in the validation of the logistic regression model. Newson (2010) and Schmid et al. (2016) report that Harrel's $C \in [0,1]$ statistic measures the predictive power of a regression model, and that C=1 corresponds to a perfect relationship between the response and the covariates.

2.4.5 Interpreting the fitted logistic regression model

According to Dayton (1992), we may interpret the results from a logistic regression model in three different levels. Firstly, each term in the equation represents contributions to estimated log-odds. Thus, for each one unit increase or decrease in X_j, there is prediction of an increase or decrease of β_j units in the log-odds in favour of $Y = 1$. Also, if all predictors are set equal to 0, the predicted log-odds in favour of $Y = 1$ would be a constant β_0. Secondly, since most people do not find it natural to think in terms of log-odds, the logistic regression equation can be transformed to odds by exponentiation as in model (31):

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = e^{\hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}}. \qquad (39)$$

Finally, the results can be expressed in terms of probabilities using the logistic function. The interpretation of the coefficients $(e^{\hat{\beta}_j})$ parallels that of the hazards in (section 2.3.4), that is for continuous covariates, if $\hat{\beta}_j > 0$ then a unit increase in the j^{th} covariate increases the hazard by $(OR_j - 1)100\%$. Otherwise, a unit increase in the j^{th} covariate decreases the hazard by $(OR_j - 1)100\%$. For categorical variables, the j^{th} covariate is actually the j^{th} category/level of the categorical variable. Hence, if $\hat{\beta}_j > 0$ then patients in the j^{th} category/level face a hazard $(OR_j - 1)100\%$ greater than those in the specified reference category/level. Otherwise, the patients in the j^{th} category/level face a hazard $(OR_j - 1)100\%$ lower than those in the specified reference category/level. In this study, OR=1, implies no association between the covariate(s)

and death, OR<1 means covariates are protective or do not favour death while OR>1 implies that death is associated with the covariate(s).

2.5 Quantile regression in survival analysis

Recall that $F(t) = P(T \leq t) = 1 - S(t)$ (equation (1)) is the distribution function of the random variable T representing the survival time of a HIV+ terminal patient during the observation period. Then for $\alpha \in (0,1)$, $Q(\alpha) = F^{-1}(\alpha)$ is the quantile function of the random variable T. For example, when $\alpha = 0.5$, then $Q(0.5) = F^{-1}(0.5)$ is the median of the distribution of T. $Q(\alpha)$ for all $\alpha \in (0,1)$ provides a complete picture of the distribution of T. Suppose that the potential factors which affect the survival of HIV+ patients are represented as in (section 2.2.1), and that they are related to the survival time by the regression model:

$$T_i = exp\{\boldsymbol{X}_i^T\boldsymbol{\beta} + \sigma\varepsilon_i\}, \tag{40}$$

where T_i is the survival time of the i^{th} patient, σ is an unknown scale parameter and the ε_i are independent random errors with an assumed distribution. Model (40) is called an accelerated failure time (AFT) model, from which the quantile regression model corresponding to the Cox regression model in (section 2.2.2) is (Chaudhuri et al., 1997; Peng et al., 2008; Flom & Peter Flom Consulting, 2011):

$$Q_i(\alpha, \boldsymbol{\beta}^\alpha) = exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}^\alpha\} = inf\{t: P(T_i \leq t|X_i^T) \geq \alpha\} = F_i^{-1}(\alpha), \tag{41}$$

where $F_i(t)$ is the distribution function of T_i which is derived from the distribution of the ε_i, and $\boldsymbol{\beta}^\alpha$ is the unknown vector of effects of the covariates on the α^{th} quantile of the distribution of T_i. $\boldsymbol{\beta}^\alpha$ may vary with α which means $\boldsymbol{\beta}^\alpha$ measures the effect of covariates not only in the centre but also in the upper and lower tails of the survival time distribution.

Quantile regression was introduced by Koenker and Basset (1978) and has become an increasingly important tool in statistical analysis (Leng & Tong, 2013). Gorfine et al. (2017), reported that quantile regression provides a framework for modelling the relationship between

an outcome and covariates using conditional quantile functions. Quantile regression is becoming an attractive alternative to the Cox (1972) proportional hazards and to the AFT models (Gorfine et al., 2014). Quantile regression models are considered robust and flexible in the sense that they can capture a variety of effects at different quantiles of the survival distribution (Gorfine et al., 2014). Lin et al. (2013) reported that quantile regression is a direct and flexible approach to modelling survival times without the proportionality hazard constraint of the Cox model. Furthermore, the authors reported that in clinical studies, quantile regression is helpful for identifying and distinguishing important prognostic factors for patient subpopulations that are characterised by short and long survival times. Lastly, the authors reported that quantile regression model has two advantages, which are: there is no need to specify a parametric form of the lifetime distribution and the regression model can be fit for each quantile. Bellavia (2015) presented the advantages of quantile regression as: firstly, the regression allows focusing on specific quantiles of interest; secondly, the entire shape of the distribution is considered; and lastly, the quantile regression is the common way to model quantiles. In addition, quantile regression does not assume any particular parametric distribution and it does not assume constant variance for the response variable as in the case of least squares regression (Rodriguez, 2017).

2.5.1 Inferences about the quantile regression model parameters

The estimate of $\boldsymbol{\beta}^{\alpha}$ is (Yu & Moyeed, 2001):

$$\widehat{\boldsymbol{\beta}^{\alpha}} = \min_{\boldsymbol{\beta}^{\alpha}} \sum_{i=1}^{n} \left| ln\, T_i - \boldsymbol{X}_i^T \boldsymbol{\beta}^{\alpha} \right| + (2\alpha - 1)(\, ln\, T_i - \boldsymbol{X}_i^T \boldsymbol{\beta}^{\alpha}), \tag{42}$$

assuming there are no censored times. The quantile regression model for known censored quantile regression (CQR) was introduced by Powell (1984) and allows semi-parametric estimation of quantile regressions in a robust way (Zhu et al., 2012). Using the CQR model, the estimate of β^{α} is (Blundell, 2007; Maposa, 2016):

$$\widetilde{\boldsymbol{\beta}^{\alpha}} = \min_{\boldsymbol{\beta}^{\alpha}} \sum_{i=1}^{n} \left| T_i^* - \min_{\boldsymbol{\beta}^{\alpha}} (\boldsymbol{X}_i^T \boldsymbol{\beta}^{\alpha}, C_i^*) \right| + (2\alpha - 1)(\, T_i^* - \min_{\boldsymbol{\beta}^{\alpha}} (\boldsymbol{X}_i^T \boldsymbol{\beta}^{\alpha}, C_i^*)), \tag{43}$$

where $T_i^* = \min_{\boldsymbol{\beta}^\alpha}(\ln T_i, \ln C_i)$, C_i is the observed right censored time, and $C_i^* = \ln C_i$. On the other hand, if the censored times (C) are random and independent of the uncensored times (T), then the estimate of $\boldsymbol{\beta}^\alpha$ is $\overline{\boldsymbol{\beta}^\alpha}$ which solves the estimating equation (Leng & Tong, 2013):

$$M_n\left(\overline{\boldsymbol{\beta}^\alpha}\right) = \sum_{i=1}^n \boldsymbol{X}_i \left[\frac{I(T_i^* - \boldsymbol{X}_i^T\overline{\boldsymbol{\beta}^\alpha} \geq 0)}{\bar{S}_{KM}(\boldsymbol{X}_i^T\overline{\boldsymbol{\beta}^\alpha})} - (1-\alpha)\right] \approx 0, \tag{44}$$

where $\bar{S}_{KM}(\boldsymbol{X}_i^T\overline{\boldsymbol{\beta}^\alpha})$ is the Kaplan-Meier estimate of the conditional survival function of the distribution of C given the covariates $\boldsymbol{X}_i$. The initial estimate of $\overline{\boldsymbol{\beta}^\alpha}$ for iteratively solving equation (44) is found by solving the equation (Leng & Tong, 2013):

$$M_n\left(\overline{\boldsymbol{\beta}^\alpha}\right) = \sum_{i=1}^n \frac{I(T_i^* \leq C_i^*)}{\bar{S}_{KM}(T_i^*)} \boldsymbol{X}_i \left[I(T_i^* - \boldsymbol{X}_i^T\overline{\boldsymbol{\beta}^\alpha} \geq 0) - (1-\alpha)\right] \approx 0. \tag{45}$$

Other authors use the Nelson-Aalen estimate in place of the Kaplan-Meier estimate (Peng & Huang, 2008; Leng & Tong, 2013).

Under the regularity conditions stated in Leng and Tong (2013), the estimator $\overline{\boldsymbol{\beta}^\alpha}$ is consistent and has an asymptotic normal distribution with covariance matrix, which depends on the unknown conditional density function of $\boldsymbol{C}$ given the covariates, in theorem 2 of (Leng & Tong, 2013). The authors suggest estimating the covariance matrix using the bootstrap method with many bootstrap replications especially if the sample size is finite. Both estimators $\widetilde{\boldsymbol{\beta}^\alpha}$ and $\overline{\boldsymbol{\beta}^\alpha}$ have been proven by the authors who developed them to be asymptotically normal. Hence it is intuitive that if the sample is large, inferences about $\boldsymbol{\beta}^\alpha$ and its components, the β_j^α $(j = 1,2,\dots,p)$, can be done using the Wald test statistic, (see Section 2.3.2),

$$\chi = \left(\widetilde{\boldsymbol{\beta}^\alpha} \text{ or } \overline{\boldsymbol{\beta}^\alpha} - \boldsymbol{\beta}^\alpha\right)^T \left[\widehat{\boldsymbol{Cov}}(\widetilde{\boldsymbol{\beta}^\alpha} \text{ or } \overline{\boldsymbol{\beta}^\alpha})\right]^{-1} \left(\widetilde{\boldsymbol{\beta}^\alpha} \text{ or } \overline{\boldsymbol{\beta}^\alpha} - \boldsymbol{\beta}^\alpha\right) \sim \chi_p^2. \tag{46}$$

$$z = \frac{\widetilde{\beta_j^\alpha} \text{ or } \overline{\beta_j^\alpha} - \beta_j^\alpha}{se(\widetilde{\beta_j^\alpha} \text{ or } \overline{\beta_j^\alpha})} \sim N(0,1), j = 1,2,\dots,p \text{ (approximately)}, \tag{47}$$

where $\widehat{\boldsymbol{Cov}}(.)$ is a "good" estimate of the covariance matrix and $se(.)$ is the $(jj)^{th}$ element of $\widehat{\boldsymbol{Cov}}(.)$. Otherwise for finite sample samples inferences can be done using the bootstrap estimates of the empirical distributions of $\widehat{\boldsymbol{\beta}^{\alpha}}$ or $\overline{\boldsymbol{\beta}^{\alpha}}$ as reported in (Leng & Tong, 2013).

2.5.2 Model selection and diagnosis in quantile regression models

The main objective is to select important predictors that have nonzero effect on the α^{th} conditional quantile of the quantile regression model:

$$ln\, T_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^{\alpha} + \epsilon_i^{\alpha}, \tag{48}$$

using redistribution-of-mass method (Wang et al., 2013). The redistribution-of-mass method entails redistribution of probability of masses $P(T_i > C_i|\boldsymbol{X}_i^T)$ of censored cases to observations on the right. The conditional probability $\pi_{0i} = P(T < C_i|\boldsymbol{X}_i^T)$ is that the survival time of the i^{th} patient is not censored. If known, then $\boldsymbol{\beta}^{\alpha}$ can be estimated by minimizing the following weighted quantile objective function with respect to $\boldsymbol{\beta}^{\alpha}$ (Wang et al., 2013):

$$L(\boldsymbol{\beta}^{\alpha}, w_0) = \sum_{i=1}^{n}\{w_{0i}\rho_{\alpha}(ln\, T_i - \boldsymbol{X}_i^T\boldsymbol{\beta}^{\alpha}) + (1 - w_{0i})\rho_{\alpha}(Y^{+\infty} - \boldsymbol{X}_i^T\boldsymbol{\beta}^{\alpha})\}, \tag{49}$$

where $\rho_{\alpha}(u) = u(\{\alpha - I(u < 0)\}, Y^{+\infty}$ is any value sufficiently large to exceed $\boldsymbol{X}_i^T\boldsymbol{\beta}^{\alpha}$ for all i, and

$w_{0i} = \begin{cases} 1, & \delta_i = 1, \\ 0, & \delta_i = 0\ and\ \pi_{0i} > \alpha, \\ \frac{\alpha - \pi_{0i}}{1 - \pi_{0i}} & \delta_i = 0\ and\ \pi_{0i} \leq \alpha. \end{cases}$ $\delta_i = I(T_i \leq C_i)$ is the censoring indicator.

The sub gradient of the weighted objective function (49) is given by,

$$nM_n(\boldsymbol{\beta}^{\alpha}, w_0) = \sum_{i=1}^{n} w_i\{1 - w_{0i}I(ln\, T_i - \boldsymbol{X}_i^T\boldsymbol{\beta}^{\alpha} \geq 0)\}, \tag{50}$$

and is an unbiased estimating function of $\boldsymbol{\beta}^{\alpha}$ (Wang et al., 2013). Hence, minimizing $L(\boldsymbol{\beta}^{\alpha}, w_0)$ with respect to $\boldsymbol{\beta}^{\alpha}$ leads to a consistent estimator of $\boldsymbol{\beta}^{\alpha}$. To select variables, (Wang et al., 2013) considers minimizing the penalized objective function:

$$L_{ALL}(\boldsymbol{\beta}^{\alpha}, w_0) = L(\boldsymbol{\beta}^{\alpha}, w_0) + \lambda_n \sum_{j=1}^{p} v_j \left|\beta_j^{\alpha}\right|, \quad (51)$$

where λ_n is a positive penalization parameter and v_j are adaptive weights. This approach leads to sparse coefficient estimation, and thus provides a convenient way to conduct model fitting and variable selection simultaneously.

A natural approach to checking goodness of fit of a quantile regression model of the form

$$Q_i^* = ln\, Q_i(\alpha, \boldsymbol{\beta}^{\alpha}) = \boldsymbol{X}_i^T \boldsymbol{\beta}^{\alpha}, \quad (52)$$

is to use martingale residuals and their transformations as follows (Peng & Huang, 2008). Consider a simple class of stochastic processes

$$K_n(\alpha) = n^{-1/2} \sum_{i=1}^{n} q(\boldsymbol{X}_i) M_i(\alpha; \widetilde{\boldsymbol{\beta}^{\alpha}}), \quad (53)$$

where $q(.)$ is a known bounded function and $M_i(\alpha; \widetilde{\boldsymbol{\beta}^{\alpha}}) = N_i\left(exp\{\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}^{\alpha}}\}\right) - \int_0^{\alpha} I[Z_i \geq exp\{\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}^{u}}\} dH(u)$, where H(u) = -log(1-u) and u ∈ [0,1). Now, assuming that model (53) is specified correctly, $K_n(\alpha)$ converges weakly to a mean 0 Gaussian process, the distribution of which can be approximated as follows (Peng & Huang, 2008):

$$K^*(\alpha) = n^{-1/2} \sum_{i=1}^{n} q(\boldsymbol{X}_i) M_i(\alpha; \widetilde{\boldsymbol{\beta}^{\alpha}})(1 - \xi_i) + n^{-1/2} \sum_{i=1}^{n} q(\boldsymbol{X}_i)\{M_i(\alpha; \boldsymbol{\beta}^{\alpha}) - M_i(\alpha; \widetilde{\boldsymbol{\beta}^{\alpha}})\}$$

(54)

Now, consider the $\xi_i's$ as random and $\{Z_i, \delta_i, X_i\}_{i=1}^{n}$ as fixed in $K^*(\alpha)$. To approximate the

null distribution of $K^*(\alpha)$, a number of realizations are simulated from $K^*(\alpha)$ by repeatedly generating $\{\xi_i\}_{i=1}^n$. Finally, an unusual pattern of $K(\alpha)$ compared with $K^*(\alpha)$ in the plot of $K(\alpha)$ with a few realizations of $K^*(\alpha)$, suggests a lack of fit of model (54).

In quantile regression models, Koenker and Machado (1999) proposed the following goodness of fit measure similar to R^2 in linear regression:

$$R^1(\alpha) = \frac{min \sum_i \rho_\alpha(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^{\alpha} - Q_\alpha(ln\, T))}{min \sum_i \rho_\alpha(y_i - Q_\alpha(ln\, T))} = 1 - \frac{min \sum_i \rho_\alpha(ln\, T_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^{\alpha})}{min \sum_i \rho_\alpha(ln\, T_i - Q_\alpha(ln\, T))}, ln\, T = \boldsymbol{X}^T \boldsymbol{\beta}^{\alpha} + \sigma\varepsilon, \quad (55)$$

where $Q_\alpha(\ln T)$ denotes the unconditional α^{th} quantile of the distribution of $\ln T$, and the minimum is with respect to the effect of covariate X_k on the α^{th} quantile of the distribution of survival time T. Like R^2, the value of $\mathrm{R}^1(\alpha)$ lies between 0 and 1. Unlike R^2 which is a global measure of goodness of fit, $\mathrm{R}^1(\alpha)$ measures the relative success of the corresponding quantile regression model and can thus be interpreted as a local goodness of fit value for a particular quantile (Schulze, 2004).

2.5.3 Interpretation of the coefficients

In quantile regression, the vector of regression coefficients, $\boldsymbol{\beta}^{\alpha}$, represents the effects of covariates on the α^{th} quantile of the distribution of the survival time T or $\ln T$, and may change with α. Therefore, the effects of a covariate in both the Cox and logistic regression models are global whereas in the quantile regression model the effect is localized on the α^{th} quantile of the distribution of the survival time T or $\ln T$. However, a coefficient with value of zero implies no association between the covariate and α^{th} quantile of the distribution of the survival time T or $\ln T$, a positive coefficient means the covariate is protective or favours survival time for a particular α while a negative coefficient implies that survival time is negatively associated with the covariate for a particular α. In other words and mathematically, the effect of covariate X_k on the α^{th} quantile of the distribution of survival time T or $\ln T$ is the change in the quantile due to a change in X_k (ceteris paribus) is (Schulze, 2004):

$$\beta_k^{\alpha} = \frac{\partial Q_\alpha(ln\, T)}{\partial_{X_k}}, ln\, T = \boldsymbol{X}^T \boldsymbol{\beta}^{\alpha} + \sigma\varepsilon \quad (56)$$

or

$$\frac{\partial Q_\alpha(T)}{\partial_{X_k}} = \beta_k^\alpha e^{\boldsymbol{X}^T\boldsymbol{\beta}^\alpha}, T = e^{\boldsymbol{X}^T\boldsymbol{\beta}^\alpha} + \sigma\varepsilon. \quad (57)$$

As an example, a positive coefficient for CD4 would mean that, as CD4 increases so does survival time, while a negative coefficient for say viral load would mean that as viral load increases, survival time decreases.

2.6 Summary

Chapter two reviewed literature on definitions and assumptions of survival analysis methods. Details on survivor and hazard functions, Kaplan-Meier estimator were covered while observation window and censorship were summarily described. Comparison of survivor functions using log-rank test was covered in detail. The Cox model, hazard rate and maximum likelihood estimates were covered with the aid of equations. Cox model development was described under the following: theoretical aspects of the model, application of the model, inference about model parameters, and model selection. Parametric models, as an alternative to Cox model were briefly explained. Cox model diagnostics was covered with some attention on proportionality and linearity assumptions, outliers and influential observations, interpretation of the coefficients of the model, goodness of the model and validation of the model. Logistic regression was included in this chapter and it considers the association between a dichotomous variable (mortality) and independent variables. Lastly, quantile regression is presented in this chapter to model the relationship between quantiles of survival time and the covariates.

CHAPTER 3

DATA ANALYSIS AND RESULTS

INTRODUCTION

Chapter 2 reviewed survival data analysis methods which are applied in this chapter to fulfil the objectives of this study as described in chapter 1. It is in this chapter where the data from the two hospitals in the Albert Luthuli municipality of the Gert Sibande district of Mpumalanga province is analysed. Data analysis methods used in this chapter, but not reviewed in chapter 2, are briefly explained and references given. The main statistical software package that was used to analyse the data is SAS Version 9.4. The other statistical packages that were used for data analysis are Stata Version 15 and R Version 3.5.1.

3.1 Preliminary data analysis

Recall from chapter 1 that the study extracted a sample of 357 from the total of 719 HIV+ terminal patients. The sample was observed over a period of 7.5 years from the 1st of January 2010 to the 30th of June 2017 at two hospitals. This study examines the effects of the covariates presented in Table 1.6.5 (in Section 1.6.5) on the survival times (and probabilities) of HIV+ terminal patients following the initiation of ART.

Table 3.1.1: Frequency distribution of Patient status by Hospital on last follow-up visit

Frequency		**Patient status**				
%	Hospital	Transferred out of hospital	Lost to follow-up	Dead	Still alive	Total
Row %	Carolina	38	17	4	16	**75**
Column %		10.6	4.8	1.1	4.5	**21.0**
		50.7	22.7	5.3	21.3	
		21.6	18.3	9.8	34.0	
	Embhuleni	138	76	37	31	**282**
		38.7	21.3	10.4	8.7	**79.0**
		48.9	27.0	13.1	11.0	
		78.4	81.7	90.2	66.0	
	Total	**176**	**93**	**41**	**47**	**357**
	Percentage	**49.3**	**26.0**	**11.5**	**13.2**	**100.0**

The follow-up time for all patients started at the time the patient got initiated into the ART programme at the hospitals' wellness centres. The main event of interest in this research was 'the death of a HIV+ terminal patient.'

Table 3.1.1 shows that by the end of the observation period 11.5% of the patients were dead with Embuleni hospital having the higher mortality rate (10.4%). The 11.5% overall mortality indicates a censoring rate (88.5%) which is too high. This questions the appropriateness of the classical survival analysis methods for analyzing the data in this study. Finally, there is an association between Patient status and Hospital (chi-square test *p-value*=0.0385).

The cross table of Patient status by Gender (Table 3.1.2) show that male have a higher overall mortality rate (5.9%) than females (5.6%) at the end of the observation period. The difference of the mortality rates is small which means the survival times were almost equally highly censored. Patient status was not associated with Gender (chi-square test *p-value*=0.3566).

Table 3.1.2: Frequency distribution of Patient status by Gender on last follow-up visit

Frequency		Patient status				
%	Gender	Transferred out of hospital	Lost to follow-up	Dead	Still alive	Total
Row %	Males	62	39	20	20	**141**
Column		17.4	11.0	5.6	5.6	**39.5**
		44.0	27.7	14.2	14.2	
		35.2	42.0	48.8	42.6	
	Females	114	54	21	27	**216**
		32.0	15.1	5.9	7.6	**60.5**
		52.8	25.0	9.7	12.5	
		64.8	58.1	51.2	57.5	
	Total	**176**	**93**	**41**	**47**	**357**
	Percentage	**49.3**	**26.0**	**11.5**	**13.2**	**100.00**

Table 3.1.3 shows that the overall mortality rate was highest among adults (10.4%) compared to that of the combined children and adolescent groups (1.1%). This means the censoring rate was too high in the children and adolescent groups. Although there was no association between Patient status and Age group (chi-square *p-value*=0.3626) this conclusion is unreliable due some cells in Table 3.1.3 having less than five counts. Again, this questions the appropriateness

of the classical survival analysis methods for analyzing the data in this study as was also mentioned above.

Table 3.1.3: Frequency distribution of the Patient status by Age group on last follow-up visit

Frequency		Patient status				
%	Age group	Transferred out of hospital	Lost to follow-up	Dead	Still alive	Total
Row %	Children	12	3	1	4	**20**
Column		3.4	0.8	0.3	1.1	**5.6**
		60	15	5	20	
		6.8	3.2	2.4	8.5	
	Adolescents	14	14	3	5	**36**
		3.9	3.9	0.8	1.4	**10.1**
		38.9	38.9	8.3	13.9	
		8.0	15.0	7.3	10.6	
	Adults	150	76	37	38	**301**
		42.0	21.3	10.4	10.6	**84.3**
		49.8	25.3	12.3	12.6	
		85.2	81.7	90.2	80.9	
	Total	**176**	**93**	**41**	**47**	**357**
	Percentage	**49.3**	**26.0**	**11.5**	**13.2**	**100.0**

The above cross tabulation results prompted the decision to regard the survival times of patients who were lost to follow-up or transferred out of the hospitals as well as of those patients who were still alive at the end of the observation period as right censored times. The rationale of doing this, which concurs with survival analysis assumptions in chapter 2, was also that the survival times of patients who were lost to follow-up or transferred out of hospitals were recorded up to the time they were lost or transferred.

Thus, in the data the Patient status became 0=Transferred or Lost to follow-up or Still alive (or censored) and 1=Dead (or uncensored). This classification of patients prompted the question of whether there was an association between Patient status and the covariates in Table 1.6.5 (section 1.6.5). If so, those covariates would be expected to also have an effect on the survival times (and probabilities) of patients. To answer the question, a logistic regression model with binary response Patient status was fitted to the data using the variable selection process described in (section 2.3.3.1) of chapter 2. The null hypothesis test in Table 3.1.4 shows that the global fit of the logistic regression model is good ($p - value < 0.0003$).

Table 3.1.4: Results of fitting the logistic regression model with binary response Patient status (0=Transferred out of hospital or Lost to follow-up or Still alive and 1=Dead)

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	62.6878	16	<.0001
Score	47.7722	16	<.0001
Wald	42.9628	16	<0.0003

Joint Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment (Regimen 1)	4	11.1022	0.0254
Follow-up CD4	1	11.1458	0.0008
Follow-up mass	1	0.2835	0.5944
Marital status	3	7.1720	0.0666
Follow-up lymphocyte	1	4.6167	0.0317
ln (Baseline Viral Load)	1	8.5984	0.0034
Follow-up Alanine Transaminase	1	4.4359	0.0352
Follow-up mass*Treatment (Regimen 1)	4	11.9518	0.0177

Association of Predicted Probabilities and Observed Responses

Percent Concordant	86.8	Somers' D	0.736
Percent Discordant	13.2	Gamma	0.736
Percent Tied	0.0	Tau-a	0.150
Pairs	12956	c	0.868

The predicted probabilities and the observed responses had a high degree of association as evidenced by a high concordance value of 86.8% and a c-value of 0.868. These measures of association are discussed in (section 2.4.2.4) of chapter 2. The model has Follow-up mass by

Treatment (Regimen 1) ($p-value<0.0177$) as a significant interaction effect at the 0.1 level of significance and the following five significant main effects at the 0.1 level of significance: Follow-up CD4 ($p-value<0.0008$); Marital status ($p-value<0.0666$); Follow-up lymphocyte ($p-value<0.0317$); *ln* (Baseline viral load) ($p-value<0.0034$); and Follow-up Alanine Transaminase (*p-value*<0.0352). Follow-up mass main effects are insignificant; however, it is included in the model because it interacts significantly with Treatment (Regimen 1).

When the regression model was refitted with binary response Patient status redefined as 0 = Dead or Still alive and 1= Transferred out of hospital or Lost to follow-up, the results in Table 3.1.5 were obtained. The table also shows a good global fit of the model ($p-value<0.0001$) with one significant interaction effects as *ln* (Baseline viral load) by Age interaction effects ($p-value<0.0408$) at the 0.l level of significance, and the following significant main effects at the 0.1 level of significance: Treatment (Regimen 1) ($p-value<0.0005$); ART Adherence ($p-value<0.0037$); Follow-up haemoglobin ($p-value<0.0497$) and Baseline lymphocyte ($p-value<0.0564$). *Ln* (Baseline viral load) has insignificant effect but is included in the model because its interaction effects with age were significant. Finally, the predicted probabilities and the observed responses show high degree of association as shown by high concordance value of 74.8% and a c-value of 0.748. These results suggest that the transfer of patients out of the hospitals and/or loss of patients to follow-up was not random, hence missing observations from these patients cannot be regarded as missing at random. Furthermore, the significant effects in Tables 3.1.4 and 3.1.5 are expected to also significantly affect the survival times (and probabilities as was mentioned before) in the Cox proportional hazards model to be fitted in (section 3.2).

Table 3.1.5: Results of fitting the logistic regression model with binary response Patient status (0=Dead or Still alive and 1= Transferred out of hospital or Lost to follow-up)

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	50.7089	11	<.0001
Score	51.8648	11	<.0001
Wald	41.6613	11	<.0001

Joint Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment (Regimen 1)	4	19.9780	0.0005
ART adherence	2	11.1821	0.0037
Follow-up haemoglobin	1	3.8522	0.0497
Age	1	3.8803	0.0489
ln (Baseline viral load)	1	2.5896	0.1076
Baseline lymphocyte	1	3.6390	0.0564
Age**ln* (Baseline viral load)	1	4.1834	0.0408

Association of Predicted Probabilities and Observed Responses

Percent Concordant	74.8	Somers' D	0.495
Percent Discordant	25.2	Gamma	0.495
Percent Tied	0.0	Tau-a	0.184
Pairs	23672	c	0.748

3.2 Semiparametric and nonparametric analysis of the data

In Section 3.1, it was decided that survival times of patients who were lost to follow-up or transferred out of the hospitals as well as of those patients who were still alive at the end of the observation period, should be regarded as right censored times. The rationale being that the survival times of patients who were lost to follow-up or transferred out of hospitals were recorded up to the time they were lost or transferred, and that missing observations from these patients cannot be regarded as missing at random. This should be kept in mind in this section in which semiparametric and nonparametric survival data analysis methods are applied to fulfil the objectives of this study given in Chapter 1.

3.2.1 Cox PH regression modelling

The Cox PH regression model was fitted to the data using the variable selection processes described in (section 2.3.3.1) of chapter 2. The results are in Table 3.2.1. Assuming that the model assumptions hold, the model globally fits the data $(p - value < 0.0001)$, with significant Follow-up CD4 cell count by Treatment (Regimen 1) $(p - value < 0.0398)$ and Follow-up lymphocyte by TB history $(p - value < 0.0248)$ being the interaction effects at the 0.1 level of significance. Furthermore, the table shows that the following are the significant main effects at the 0.1 level of significance: ART adherence $(p - value < 0.0002)$; Age $(p - value < 0.0411)$; Follow-up mass $(p - value < 0.0482)$; Baseline sodium $(p - value < 0.0472)$ and *ln* (Baseline viral load) $(p - value < 0.0009)$.

Table 3.2.1: Results of fitting the Cox PH regression model (right censored times are for Transferred out of hospital or Lost to follow-up or Still alive patients)

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	134.1230	18	<.0001
Score	118.4437	18	<.0001
Wald	79.6312	18	<.0001

Joint Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment (Regimen 1)	4	23.0719	0.0001
ART adherence	2	16.9307	0.0002
TB history	1	3.1820	0.0745
Age	1	4.1727	0.0411
Follow-up mass	1	3.9037	0.0482
Follow-up lymphocyte	1	0.0634	0.8012
Baseline sodium	1	3.9391	0.0472
Follow-up CD4	1	0.0489	0.8250
ln (Baseline viral load)	1	11.1100	0.0009
Follow-up CD4*Treatment (Regimen 1)	4	10.0349	0.0398
Follow-up lymphocyte*TB history	1	5.0405	0.0248

3.2.1.1 Checks for outliers and influential observations

SAS provides many procedures that can be used to examine data for outliers. In this study outliers were identified from the original data set using SAS procedures PROC UNIVARIATE, PROC SGPLOT and PROC SGPANEL. PROC UNIVARIATE enlists extreme values, PROC SGPLOT plots box and whisker diagram while PROC SGPANEL plots a scatter diagram.

Figure 3.2.1 shows the typical box plot which was used to check the distribution of outliers in each covariate in the data set. The lines extending from both sides of the box represent a distance of 1.5 times the interquartile range from the sides of the box and the circles represent possible outliers;

that is data points more than 1.5 times the interquartile range below the first quartile or above the third quartile.

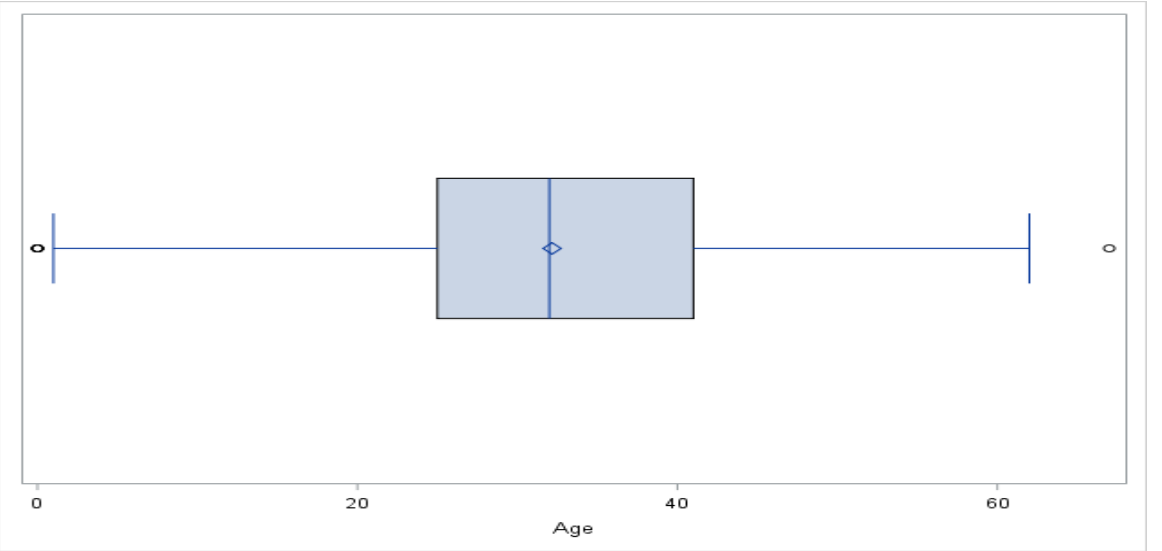


Figure 3.2.1: A typical Box plot for checking the presence of outliers in the values of a covariate

The presence of outliers in Categorical factors was checked using SAS PROC FREQ with ODS GRAPHICS OFF statements.

Table 3.2.2: A typical table for checking the presence of outliers in a Categorical covariate

Treatment levels	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NVP+D4T+3TC	88	24.7	88	24.7
EFV+D4T+3TC	175	49.0	263	73.7
EFV+AZT+3TC	26	7.3	289	81
EFV+3TC+TDF	48	13.5	337	94.4
NVP+3TC+TDF	20	5.6	357	100.0

The SAS code ‘PROC FREQ DATA=sample ORDER=freq; TABLE State Rank / MISSING;

RUN;', produces a table and a bar graph of each variable showing the percentage of subjects at each level of the variable as shown in Table 3.2.2. The presence of outliers could be evidenced by the number of rows exceeding the rows in the study design instrument, while missingness in levels would be evidenced by cumulative percent column of the table not totalling to 100. As discussed in (subsubsection 2.3.3.2), outliers were removed using PROC UNIVARIATE procedure and missingness was handled using PROC HPIMPUTE procedure for imputation to prepare data for modelling. The identification of influential observations using DFBETAS and DFFITS graphs was discussed in (subsubsection 2.3.3.2) of chapter 2. Figure 3.2.2 displays the DFBETAS graph for covariate Follow-up CD4.

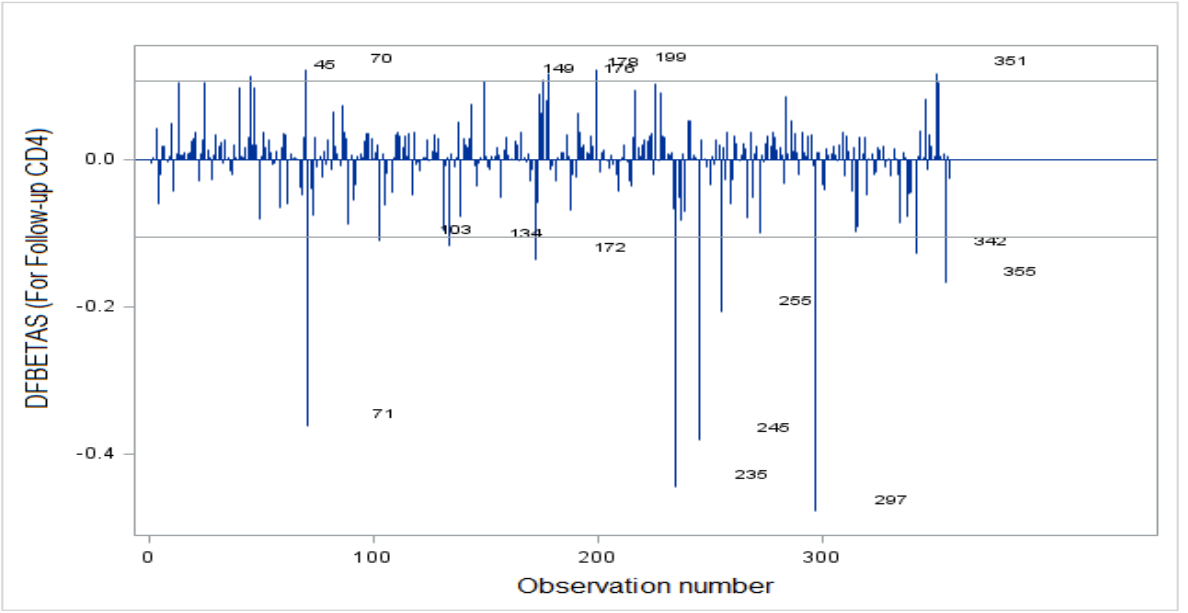


Figure 3.2.2: The DFBETAS graph for covariate Follow-up CD4 in the model in Table 3.2.1

A general cut-off to consider for both DFBETAS and DFFITS is 2; a size-adjusted cut-off recommended by Belsley, Kuh and Welsch is $2/\sqrt{n}$ for DFBETAS and $2\sqrt{p/n}$ for DFFITS where, p is the number of parameters in the model and n is the number of observations used to fit the model (SAS Institute Inc., 2015). The figure shows that there are influential observations with DFBETA values which lie outside $\pm 2/\sqrt{n} = \pm 2/\sqrt{357} = \mp 0.1057$ (two horizontal lines above and below zero). These are observation numbers 71, 235, 245, 255, 297 and 355. Similar plots for other covariates in the model (Table 3.2.1) were made, and for these plots, observation numbers 86, 177, 228, 235, 245, 297 and 338 were found to be influential with respect to more than one covariate. Although these observations are influential, the values of the covariates of these observations (displayed in Table 3.2.3) are not in error as there are within their ranges.

Furthermore, the significance/insignificance of the effects in Table 3.2.1 remained unchanged upon refitting the model to the data excluding these influential observations, hence the observations were retained in the data. As discussed in (subsection 2.3.3.3) of chapter 2, the difference in fits (DFFITS) as another measure of influence quantifies the number of standard deviations that the fitted value changes when the i^{th} observation is left out. This analysis proceeds in much the same way as in the DFBETA analysis, only that the DFFITS is a global measure of influence (that is, the DFFITS is not tied to a particular covariate).

Table 3.2.3: Influential observations with respect to more than one covariate in Table 3.2.1

Obs	Treatment (Regimen 1)	ART adherence	TB history	Age	Follow-up mass	Follow-up lymphocyte	Baseline sodium	Follow-up CD4	Ln(Baseline Viral Load)
86	6	1	2	1	17	35	135	1130	10.3
177	5	3	2	19	42	5	152	1108	4.7
182	5	2	1	42	56	23	146	317	11.5
235	5	1	2	3	15	14	139	1029	12.7
245	5	2	2	1	10	17	141	977	11.9
297	6	2	2	32	87	29	132	1048	14.0
338	6	2	2	28	72	13	122	168	4.2

Figure 3.2.3 displays the DFFITS graph for checking the presence of influential observations.

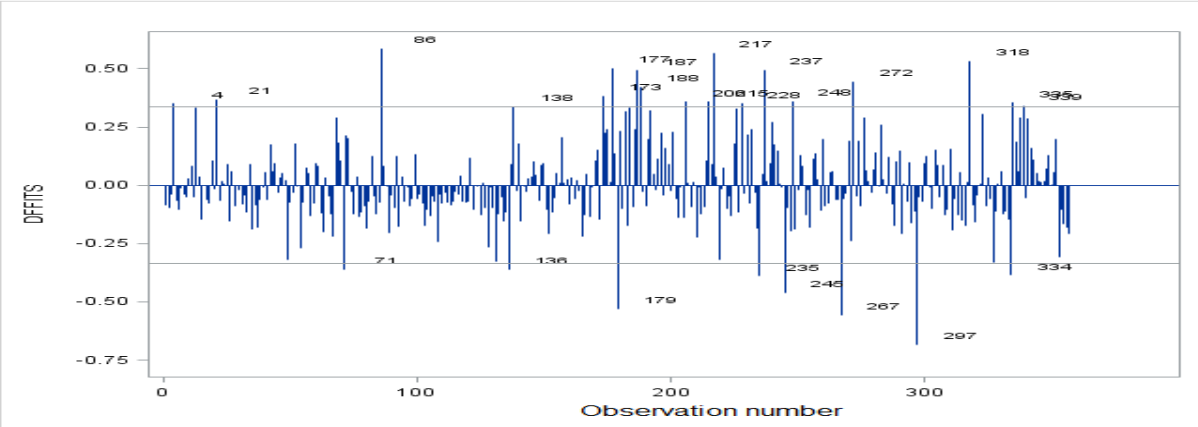


Figure 3.2.3: The DFFITS graph for checking the presence of influential observations.

As discussed in (subsection 2.3.3.3), an observation is deemed influential if the absolute value

of its DFFITS value is greater than $2\sqrt{10/357} \approx 0.3347$. The outstanding influential observations as shown in Figure 3.2.3 include 86, 177, 179, 217, 297 and 318. The DFFITS results on identified influential observations are almost the same as for DFBETAS, hence we retain the observations in the model since their covariate values are not in error and they are also within their ranges.

3.2.1.2 Checks of the linearity and proportional hazards assumptions

Table 3.2.4 shows the results of the Supremum test for the linearity of the predictor assumption, while Table 3.2.5 shows those of the Supremum test for the proportional hazards assumption after the removal of outliers from the original data set. The Supremum test was discussed in (subsection 2.3.3.2). Table 3.2.4 contains only the continuous covariates from Table 3.2.1 since linearity test is applicable on continuous variables only. The $p - values$ in Table 3.2.4 are all large (≥ 0.05) leading to the conclusion that the linearity of the predictor assumption is not violated by the data for all the covariates in the table. The same conclusion is obtained from the *p-values* of the test for the proportional hazards assumption in Table 3.2.5.

Table 3.2.4: The results of the Supremum test for the linearity of the predictor assumption

Supremum Test for Functional Form				
Variable	Maximum Absolute Value	Replications	Seed	Pr>MaxAbsVal
Age	4.3458	1000	12345	0.2280
Follow-up mass	2.8048	1000	12345	0.7540
Follow-up	2.7693	1000	12345	0.6730
Baseline sodium	2.3449	1000	12345	0.8710
Follow-up CD4	4.3598	1000	12345	0.3500
ln(Baseline viral	3.9151	1000	12345	0.6120

Table 3.2.5 contains all the covariates in the model as in Table 3.2.1, since proportional hazards assumption considers all covariates in the model. It should be noted that both graphical methods (of Schoenfield residuals) and statistical tests (Supremum tests) can be used for assessing the linearity and the PH assumptions. Statistical tests are sometimes preferred because they are more objective. This does not mean that statistical tests are better than the

graphical methods which are generally more informative but subjective (Kleinbaum & Klein, 2012).

Table 3.2.5: The results of the Supremum test for proportional hazards assumption

Supremum Test for Proportional Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr>MaxAbsVal
Treatment (Regimen 1)	0.9565	1000	12345	0.6160
ART adherence	1.2680	1000	12345	0.1170
TB history	0.7130	1000	12345	0.9440
Age	0.6991	1000	12345	0.5040
Follow-up mass	0.9879	1000	12345	0.1800
Follow-up lymphocyte	2.1781	1000	12345	0.5600
Baseline sodium	1.0577	1000	12345	0.1770
Follow-up CD4	0.6943	1000	12345	0.8960
ln (Baseline viral load)	0.4821	1000	12345	0.8340
Follow-up CD4*Treatment (Regimen	2.0708	1000	12345	0.2500
Follow-up lymphocyte*TB history	2.1743	1000	12345	0.6270

3.2.1.3 Checking for the overall goodness of fit of the model

Figure 3.2.3 displays the graph of the estimated Nelson-Aalen cumulative hazard function versus the Cox-Snell residuals (dotted line). The function approximately follows the 45° line (continuous line) which means the conditional distribution of the cumulative hazard function given the covariate vector is reasonably well approximated by the exponential distribution with hazard rate 1. This means the Cox PH model does not fit the data too badly.

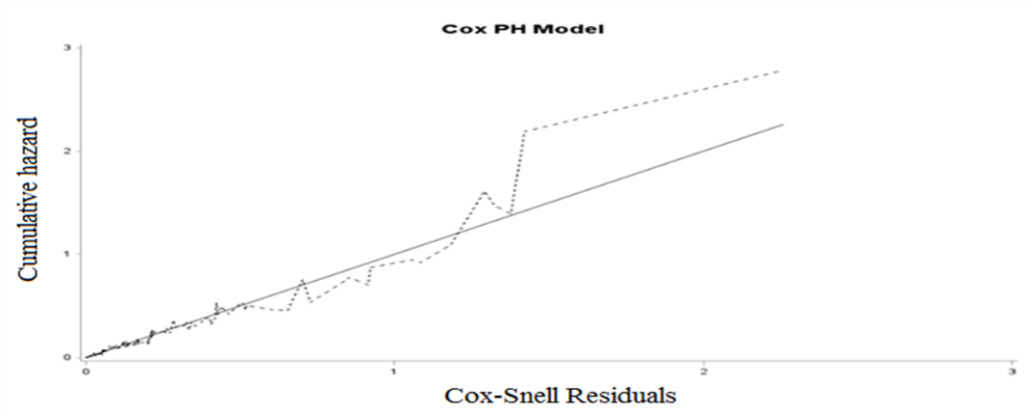


Figure 3.2.4: Estimated Nelson-Aalen cumulative hazard function versus the Cox-Snell hazard residuals (dotted line): continuous 45° theoretical line

3.2.1.4 Harrel's concordance statistic

Table 3.2.6: Estimated Harrel's concordance statistic

Harrell's concordance statistic					
Source	Estimate of C	Comparable Pairs			
		Concordance	Discordance	Tied in Predictor	Tied in Time
Model	0.9182	9080	809	0	46

As discussed in (subsection 2.4.2.4) of chapter 2, Harrel's $C \in [0,1]$ measures the predictive power of a regression model, and C=1 corresponds to a perfect relationship between the response and the covariates. The estimate of C (0.9182) from fitting the Cox PH model is in Table 3.2.6. The estimate is close to one which means the Cox PH model does not fit the data too badly as was previously concluded.

3.2.1.5 Estimated parameters of the final Cox PH model

In the context of this study, a positive parameter estimate $(\hat{\beta} > 0)$ for a covariate means the related covariate is associated with higher hazard or poorer survival chance for an HIV+ terminal patient. On the other hand, a negative parameter estimate $(\hat{\beta} < 0)$ for a covariate means the related covariate is associated with lower hazard or better survival chance. Table 3.2.7 shows that the main effects which are associated with poor survival probability at 0.1 significance level are: Age (Hazard Ratio=1.037, $p - value < 0.0411$), Baseline sodium

(Hazard Ratio=1.048, $p-value<0.0472$), *ln*(Baseline viral load) (Hazard Ratio=1.853, $p-value<0.0009$) and ART adherence (poor relative to fair ART adherence) (Hazard Ratio=5.334, $p-value<0.0004$). This means an increase in Age by one year resulted in a statistically significant increase in the HIV hazard (by about 3.7%) among the HIV+ terminal patients, while holding all other covariates constant. An increase in Baseline sodium by one-unit resulted in a statistically significant increase in the HIV hazard (by about 4.8%) among the HIV+ terminal patients, while holding all other covariates constant. Similarly; 2.718 times increase in Baseline viral load holding all other covariates constant, is associated with (about 85.3%) increase in the hazard of HIV among the HIV+ terminal patients.

The estimated 95% confidence interval of the hazard ratio for Age is (1.001, 1.073), which implies that for every one-year increase in the Age of an HIV/AIDS patient, the hazard rate is as much as 7.3% low to 0.1% lower. With regards ART adherence, patients with a poor ART adherence faced an HIV hazard that was about 18.75% greater than of those with a fair ART adherence, while holding other covariates constant. The estimated 95% confidence interval of the hazard ratio for poor ART adherence is (2.104, 13.521), which implies that the hazard ratio for poor adherence to ART relative to fair adherence to ART is as high as 13.521 and as low as 2.104. The estimated 95% confidence interval of the hazard ratio for Baseline sodium is (1.001, 1.097), which implies that for every one-unit increase in Baseline sodium of an HIV/AIDS patient, the hazard rate is as much as 9.7% low to 0.1% lower. The estimated 95% confidence interval of the hazard ratio for ln (Baseline viral load) is (1.289, 2.662), which implies that for every 2.718 times increase in Baseline viral load, the hazard rate is as much as 2.662 times low to 1.289 times lower. Table 3.2.7 also shows that the interaction effect associated with low survival probability at 0.1 significance level is Follow-up lymphocyte by TB history (yes relative to no TB history) (Hazard Ratio = 1.116, $\text{p}-\text{value}<0.0248$). This means an increase in Follow-up lymphocyte by one-unit for patients having TB history relative to those without TB history resulted in a statistically significant increase in the HIV hazard (by about 11.6%). The estimated 95% confidence interval for the interaction Follow-up lymphocyte by TB history (yes relative to no TB history), is (1.043, 1.193), which implies that the hazard ratio for this interaction is as much as 19.3% low to 4.3% lower.

The main effect which is associated with good survival probability at 0.1 significance level as in Table 3.2.7, is Follow-up mass (Hazard Ratio = 0.971, $\text{p}-\text{value}<0.0482$). A one-unit increase in Follow-up mass resulted in a statistically significant decrease in the HIV hazard (by

about 0.29%) among the HIV+ terminal patients while holding all other covariates constant. The estimated 95% confidence interval of the hazard ratio for Follow-up mass is (0.943, 1.000), which implies that for every one-unit increase in Follow-up mass of an HIV/AIDS patient, the hazard rate is as much as 5.7% low to 0% lower. It is noted that Follow-up mass has no significant effect on the survival of patients at the 0.05 level of significance as evidenced by the inclusion of 1 in the 95% confidence interval of its hazard ratio (0.943, 1.000). Table 3.2.7 also shows the interaction effect which is associated with high survival probability at 0.1 significance level as Follow-up CD4 by Treatment (Regimen 1) (Hazard Ratio = 0.936, $p-value < 0.0024$). This means an increase in Follow-up CD4 by one-unit for patients treated with (EFV+AZT+3TC) resulted in a statistically significant decrease in HIV hazard (by about 6.4%). The estimated 95% confidence interval for the interaction Follow-up CD4 by Treatment (Regimen 1) is (0.897, 0.977), which implies that the hazard ratio for this interaction is as much as 10.3% low to 2.3% lower.

Table 3.2.7: Maximum likelihood estimates of the final Cox PH model

Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence	
Treatment (Regimen 1)=1 (NVP+D4T+3TC)	4.2922	2.1974	3.8156	0.0508	.	.	.
Treatment (Regimen 1)=5 (EFV+D4T+3TC)	3.5266	2.1456	2.7017	0.1002	.	.	.
Treatment (Regimen 1)=6 (EFV+AZT+3TC)	15.4000	3.3366	21.3030	<.0001	.	.	.
Treatment (Regimen 1)=19 (NVP+TDF+3TC)	-12.3093	1916	0.0000	0.9949	.	.	.
ART adherence=1 (poor)	1.6741	0.4746	12.4422	0.0004	5.334	2.104	13.521
ART adherence=3 (good)	-1.0505	0.8088	1.6869	0.1940	0.350	0.072	1.707
TB history=1 (yes)	-1.2713	0.7127	3.1820	0.0745	.	.	.
Age	0.0360	0.0176	4.1727	0.0411	1.037	1.001	1.073
Follow-up mass	-0.0293	0.0148	3.9037	0.0482	0.971	0.943	1.000
Follow-up lymphocyte	0.0080	0.0319	0.0634	0.8012	.	.	.
Baseline sodium	0.0467	0.0235	3.9391	0.0472	1.048	1.001	1.097
Follow-up CD4	0.0010	0.0043	0.0489	0.8250	.	.	.
ln (Baseline viral load)	0.6166	0.1850	11.1100	0.0009	1.853	1.289	2.662
Follow-up CD4*TreatmentR1=1	-0.0046	0.0050	0.8221	0.3646	0.996	0.991	1.002
Follow-up CD4*TreatmentR1=5	-0.0056	0.0047	1.4015	0.2365	0.995	0.992	0.999
Follow-up CD4*TreatmentR1=6	-0.0667	0.0220	9.2277	0.0024	0.936	0.897	0.977
Follow-up CD4*TreatmentR1=19	-0.0015	4.1730	0.0000	0.9997	0.999	0.000	3563.1
Follow-up lymphocyte*TB history=1	0.1013	0.0451	5.0405	0.0248	1.116	1.043	1.193

Key: TreatmentR1=Treatment (Regimen 1); Reference levels: (18=EFV+TDF+3TC) for Treatment (Regimen 1); ‘fair’ for ART adherence and ‘no’ for TB history.

Comment 1 on Table 3.2.7: The hazard ratio and the 95% hazard ratio confidence limits columns in Table 3.2.7, for the main effects of interaction are left empty. SAS omits hazard ratio entries for terms involved in interactions as a reminder that the hazard ratios corresponding to these effects depend on other variables in the model (University of California, Statistical Consulting Group , 2019).

Comment 2 on Table 3.2.7: The 95% hazard ratio confidence limits for covariates involving Treatment (Regimen 1) (NVP+TDF+3TC) are extremely wide supposedly because of small sample size and heavy censoring (low death rate) (Fay et al, 2013).

3.2.2 Nonparametric inferences about the survivor functions

The cohort of HIV+ terminal patients in this study is not homogeneous with respect to their characteristics that may affect their survival. Hence, it will be necessary to test the equality of survivor functions among groups (strata) of patients, (see section 2.2.2) of chapter 2. Log-rank tests and the Kaplan-Meier functions presented in this section are for independent factors which were found to be statistically significant in the final Cox PH model (see Table 3.2.7). Hazard ratios for pairwise comparisons of strata are presented at the end of this section in Table 3.2.16. The interpretation of Kaplan-Meier survival functions for each main factor or for an interaction of factors assumes that values of all other factors in the final Cox PH model are kept constant.

3.2.2.1 Kaplan-Meier survival functions for ART adherence groups

Figure 3.2.5 shows that HIV/AIDS patients with poor ART adherence have the lowest survival probability (as low as around 40%) while patients with good ART adherence have the highest survival probability (at least around 95%). Patients with poor ART adherence and those with fair ART adherence experience the event (death) earlier than those with good ART adherence. The test statistics in Table 3.2.8 show that ART adherence strata in Figure 3.2.5 are statistically different ($p - value < 0.0001$). As shown in Table 3.2.16, patients with poor adherence are about 6 times more likely to die from HIV/AIDS related illness relative to patients with fair ART adherence while patients with poor adherence are about 18 times more likely to die from HIV/AIDS related illness relative to patients with good adherence.

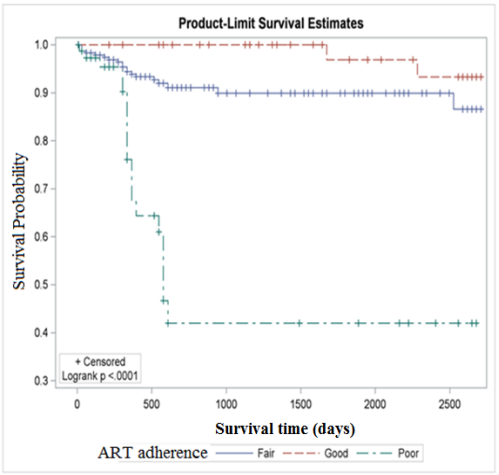


Figure 3.2.5: Kaplan-Meier survival function estimates for ART adherence strata

Table 3.2.8: Log-rank and other tests for equality of ART adherence survivor functions

Test of Equality over Strata (ART adherence)			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	52.9375	2	<.0001
Wilcoxon	41.3917	2	<.0001
-2Log(LR)*	45.0349	2	<.0001

3.2.2.2 Kaplan-Meier survival functions for Age groups

While Figure 3.2.6 shows that survivor functions of adolescent versus those of adults and children are different, the test statistics in Table 3.2.9 show the survivor functions of the three age groups are the same ($p - value \geq 0.3398$). This is further confirmed by the pairwise comparisons of the age groups using the hazard ratios in Table 3.2.16.

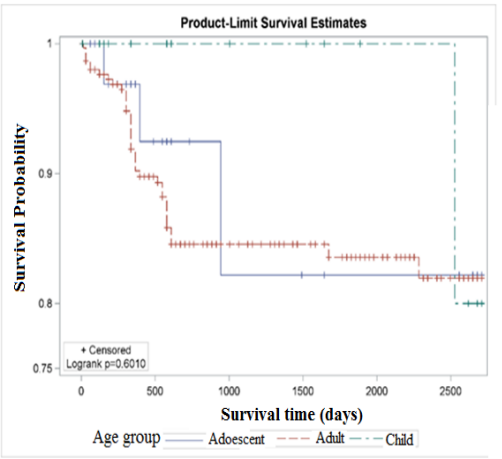


Figure 3.2.6: Kaplan-Meier survival function estimates for Age groups

Table 3.2.9: Log-rank and other tests for equality of Age group survivor functions

Test of Equality over Strata (Age)			
Test	Chi-Square	DF	Pr>Chi-Square Chi-Square
Log-Rank	1.0185	2	0.6010
Wilcoxon	2.1590	2	0.3398
-2Log(LR)*	1.4298	2	0.4892

It is noted that the conclusions from Figure 3.2.6 and the Log-rank test are questionable because of the small number of deaths (1 and 3 as in Table 1.6.3) among children and adolescents respectively due probably to sampling. The other reason is that HIV progression is not the same among children and adults, and in retrospect should have been studied separately as per recommendations in literature. Disease progression in the absence of ART is typically more rapid in HIV-infected children than in HIV-infected adults. ART can generally be

initiated earlier in children than in adults, and the decline in the viral reservoir seems to continue for longer in children (Goulder et al., 2016). In addition, AIDS typically develops more rapidly in children than in adults (Muenchhoff et al., 2016).

3.2.2.3. Kaplan-Meier survival functions for Follow-up mass

The continuous variable Follow-up mass of patients was categorized as follows (adopted from Walpole et al., 2012 & "Human body weight," 2018): Above normal: > 66kg, Normal: 63kg - 66kg, Below normal: < 63kg. As shown in Figure 3.2.7, HIV/AIDS patients with Follow-up mass below normal have the lowest survival probability (as low as 75%) while patients with normal mass have the highest survival probability (at least 95%). Patients with Follow-up mass above normal have survival probability mostly around 85% to 95%. The Log-Rank and likelihood ratio test results in Table 3.2.10 show that the survivor functions of Follow-up mass strata in Figure 3.2.7 are statistically different ($p-value \leq 0.0346$), and this is confirmed by the pairwise comparisons of the Follow-up mass strata using hazard ratios in Table 3.2.16.

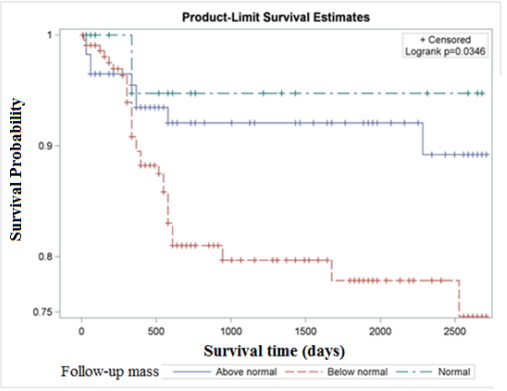


Figure 3.2.7: Kaplan-Meier survival function estimates for Follow-up mass strata

Table 3.2.10: Log-rank and other tests t for equality of Follow-up mass survivor functions

Test of Equality over Strata (Mass)			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	6.7280	2	0.0346
Wilcoxon	4.1741	2	0.1241
-2Log(LR)*	10.0556	2	0.0066

Thus, patients with Follow-up mass below normal are about twice more likely to die from HIV/AIDS related illness relative to patients with Follow-up mass above normal while the difference among other groups are hazard-wise statistically non-significant.

3.2.2.4. Kaplan-Meier survival functions for Baseline sodium

Baseline sodium was categorised as shown in the Laboratory report on Sodium in Appendix J. Both Figure 3.2.8 ($p-value \geq 0.2899$) and Table 3.2.11 ($p-value \geq 0.1369$) indicate that the survivor functions are not statistically different at 0.05 significance level. Furthermore, this

is confirmed by the pairwise comparisons of the strata using the hazard ratios in Table 3.2.16 which show no statistical difference.

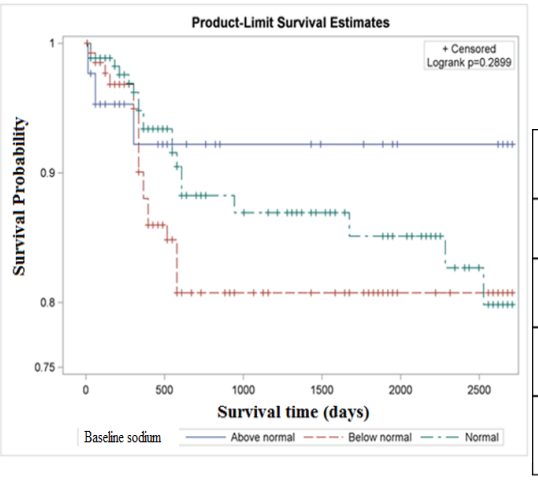


Figure 3.2.8: Kaplan-Meier survival function estimates for Baseline sodium strata

Table 3.2.11: Log-rank and other tests for equality of Baseline sodium survivor functions

Test of Equality over Strata (Sodium)			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	2.4761	2	0.2899
Wilcoxon	2.6912	2	0.2604
-2Log(LR)*	3.9765	2	0.1369

However, the visual order of the graphs in Figure 3.2.8 shows that HIV/AIDS patients with baseline sodium below normal have the lowest survival experience.

3.2.2.5. Kaplan-Meier survival functions for Baseline viral load

Figure 3.2.9 shows that HIV/AIDS patients with Baseline viral load above 5000 HIV RNA copies/mm^3 have the lowest survival probability (as low as around 65%) while patients with Baseline viral load below 50 HIV RNA copies/mm^3 (undetectable viral load) have the highest survival probability (100%) throughout the follow-up period. The test statistics in Table 3.2.12 indicate that the survivor functions of the Baseline viral load strata are statistically different ($p-value < 0.0001$). The hazard ratios in Table 3.2.16 show that patients with Baseline viral load between 50 HIV RNA copies/mm^3 and 5000 HIV RNA copies/mm^3 have HIV hazard which is around 87% lower relative to patients with Baseline viral load above 5000 HIV RNA copies/mm^3.

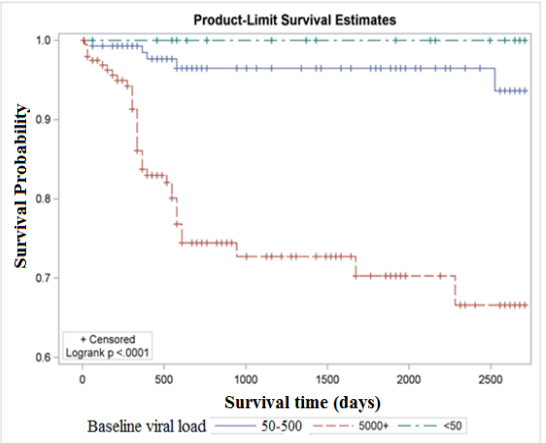


Figure 3.2.9: Kaplan-Meier survival function estimates for Baseline viral load strata

Table 3.2.12: Log-rank and other tests for equality of Baseline viral load survivor functions

Test of Equality over Strata of Baseline viral load			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	30.9448	2	<.0001
Wilcoxon	27.1307	2	<.0001
-2Log(LR)*	43.6135	2	<.0001

3.2.2.6 Kaplan-Meier survival functions for Follow-up lymphocyte by TB history groups

Patients with Follow-up lymphocyte below normal level and with TB history have the lowest survival experience, and they experienced the event (death) earlier than other groups (see Figure 3.2.10). In addition, patients belonging to this stratum did not reach the end of the study alive. Patients with normal Follow-up lymphocyte levels and without TB history have the highest survival experience. The survivor functions of the strata in Figure 3.2.10 are significantly different as evidenced by the p-values in Table 3.2.13 ($p-value < 0.0001$). The hazard ratios in Table 3.2.16 show that patients with TB history relative to those without TB history (at Follow-up lymphocyte below normal) are about 2.8 times likely to experience HIV related hazard (death), while patients with TB history relative to those without TB history (at normal Follow-up lymphocyte) are about 52 times likely to experience HIV related hazard (death). The first comparison is for two groups which are both having health problems (in form of TB and Follow-up lymphocyte below normal) while the second comparison is of a health group (TB negative and normal Follow-up lymphocyte) and a group with health problems. In addition, and as from descriptive statistics, the mortality among TB positive patients (at normal Follow-up lymphocyte) is around 67% while the mortality of TB positive patients (at Follow-up lymphocyte below normal) is around 21%.

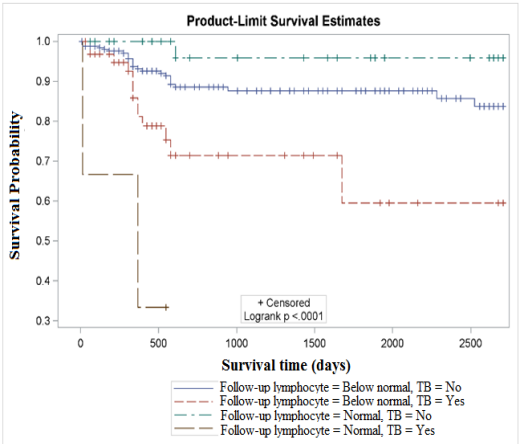


Figure 3.2.10: Kaplan-Meier survival function estimates for Follow-up lymphocyte by TB history strata

Table 3.2.13: Log-rank and other tests for equality of Follow-up lymphocyte by TB history survivor functions

Test of Equality over Strata (Follow-up lymphocyte by TB)			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	29.5848	3	<.0001
Wilcoxon	30.0687	3	<.0001
-2Log(LR)*	26.2431	3	<.0001

3.2.2.7: Kaplan-Meier survival function estimates for Follow-up CD4 by Treatment (Regimen 1)

In Figure 3.2.11, patients with Follow-up CD4 above 200 cells/mm^3 and taking (EFV+3TC+TDF) or (EFV+D4T+3TC) have the highest survival experience if other covariates are kept constant. On the other hand, patients with Follow-up CD4 below 200 cells/mm^3 and taking (EFV+AZT+3TC) have the lowest survival experience, if other covariates are kept constant. The difference across the Follow-up CD4 by Treatment strata are statistically significant (Table 3.2.14, $p-values < 0.0001$).

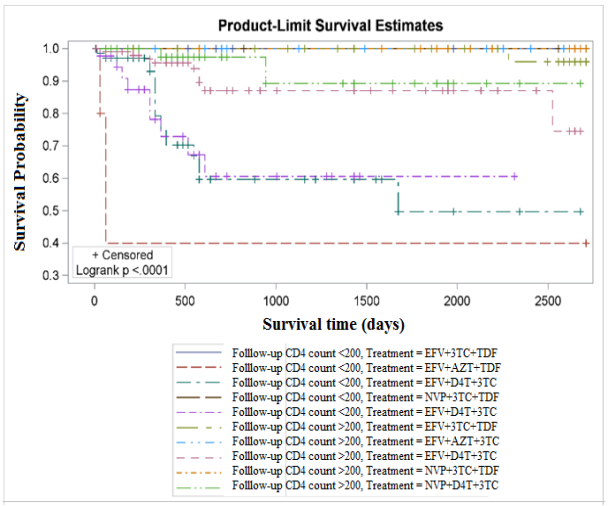


Figure 3.2.11: Kaplan-Meier survival function estimates for Follow-up CD4 by Treatment

Table 3.2.14: Log-rank and other tests for equality of Follow-up CD4 count by Treatment (Regimen 1) survivor functions

Test of Equality over Strata (Follow-up CD4 by Treatment)			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	69.6268	9	<.0001
Wilcoxon	76.1848	9	<.0001
-2Log(LR)*	74.4060	9	<.0001

3.2.2.8 Kaplan-Meier survival functions for Follow-up CD4 and related results

CD4 cell count is a key independent factor in this study because it remains the best measurement of a patient's immune and clinical status, and supports diagnostic decision-making, particularly for patients with advanced HIV disease. Figure 3.2.12 shows some relevant strata in CD4 count while Figure 3.2.13 shows the bar graphs for Baseline CD4 and Follow-up CD4.

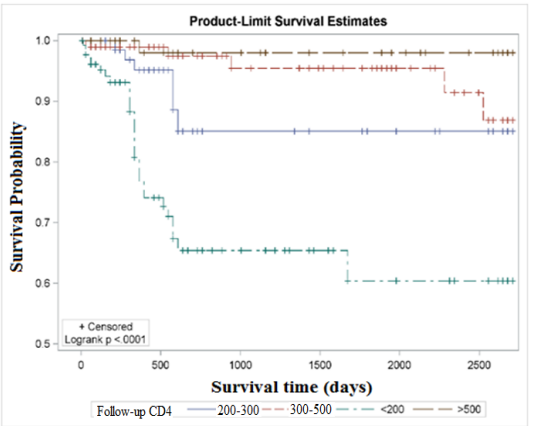


Figure 3.2.12: Kaplan-Meier survival function estimates for Follow-up CD4 Count

Table 3.2.15: Log-rank and other tests for equality of Follow-up CD4 strata survivor functions

Test of Equality over Strata (Follow-up CD4)			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	45.4274	3	<.0001
Wilcoxon	44.8056	3	<.0001
-2Log(LR)	51.3597	3	<.0001

The Kaplan-Meier survival functions for Follow-up CD4 count (figure 3.2.12) show that patients with: Follow-up CD4 count above 500 cells/mm^3, Follow-up CD4 between 200 cells/mm^3 and 500 cells/mm^3 and Follow-up CD4 count below 200 cells/mm^3 have comparatively high, moderate and low survival experiences respectively, assuming that other covariates are held constant. The difference across the Follow-up CD4 strata are statistically significant (Log-Rank, $p - value < 0.0001$). The big vertical gaps separating graphs confirm the difference across the Follow-up CD4 strata. As shown in Figure 3.2.13, following ART initiation; out of the 357 patients, the Baseline CD4 counts were as follows: 254 (71.14%) had CD4 less than 200 cells/mm^3, 80 (22.41%) had CD4 between 200 cells/mm^3 and 500 cells/mm^3 while 23 (6.44%) had CD4 greater than 500 cells/mm^3. On the other hand, the follow-up CD4 counts were as follows: 139 (38.94%) had CD4 less than 200 cells/mm^3, 165 (46.22%) had CD4 between 200 cells/mm^3 and 500 cells/mm^3 while 53 (14.85%) had CD4 greater than 500 cells/mm^3. The condition of HIV disease when CD4 is below 200 cells/mm^3 is called AIDS.

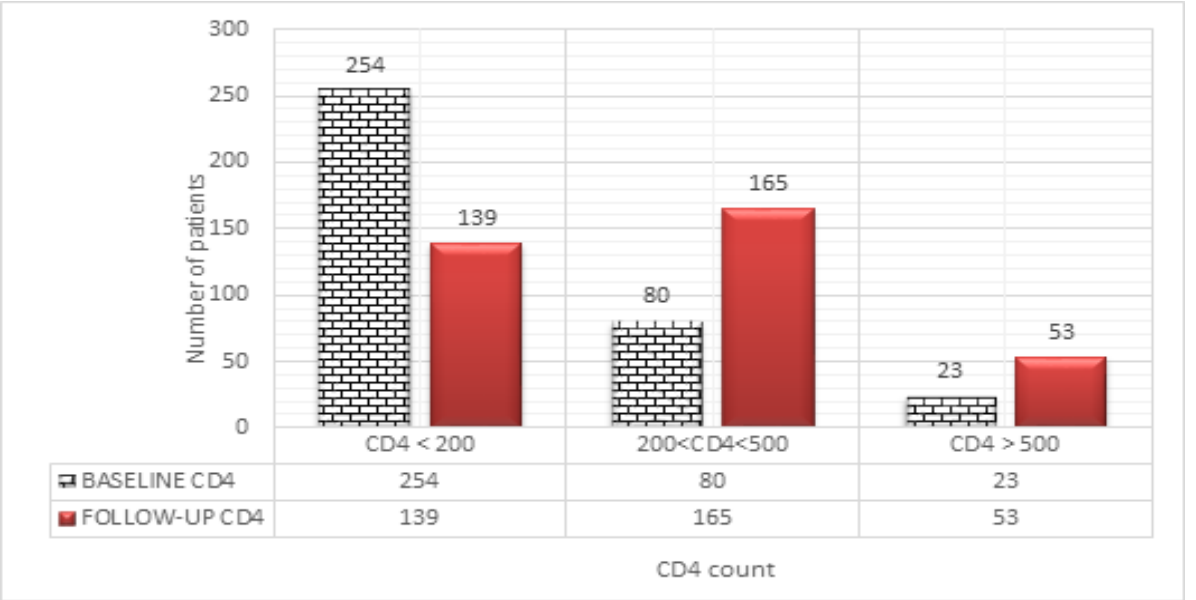


Figure 3.2.13: Bar graphs for the distribution of Baseline and Follow-up CD4 in HIV/AIDS patients

Table 3.2.16: Hazard ratios for the main and interaction effects strata

Factor	**Point Estimate**	**95% Wald Confidence Limits**	
TB history (yes vs no) at Follow-up lymphocyte =Below normal	2.847	1.445	5.608
TB history (yes vs no) at Follow-up lymphocyte=Normal	51.832	4.615	582.112
Treatment(Regimen1) (NVP+D4T+3TC vs EFV+3TC+TDF)	14.110	1.780	111.862
Treatment(Regimen1) (EFV+D4T+3TC vs EFV+3TC+TDF)	15.498	2.067	116.200
Treatment(Regimen1) (NVP+D4T+3TC **vs** EFV+D4T+3TC)	0.910	0.451	1.838
Treatment(Regime1) (NVP+D4T+3TC **vs** EFV+AZT+3TC)	2.941	0.636	13.595
Treatment(Regimen1) (EFV+D4T+3TC **vs** EFV+AZT+3TC)	3.230	0.753	13.851
Treatment(Regimen1) (EFV+AZT+3TC **vs** EFV+3TC+TDF)	4.798	0.435	52.932
TB history (yes vs no)	3.523	1.848	6.716
ART adherence (poor vs fair)	6.039	3.198	11.405
ART adherence (poor vs good)	18.441	4.233	80.332
Follow-up mass (Above normal vs Below normal)	0.433	0.205	0.912
Baseline viral load (viral load 50-5000 vs viral load	0.130	0.051	0.333
Follow-up CD4 (CD4 count <200 vs CD4 count >200)	5.493	2.731	11.048
Follow-up CD4 (CD4 count <200 vs CD4 count >500)	25.826	3.496	190.766
Follow-up lymphocyte (Below normal vs Normal)	1.810	0.557	5.878
Baseline sodium (Above normal vs Below normal)	0.443	0.131	1.499
Baseline sodium (Above normal vs Normal)	0.640	0.189	2.164
Baseline sodium (Below normal vs Normal)	1.445	0.765	2.731
Age group (Child vs Adolescent)	0.518	0.054	4.984
Age group (Child vs Adult)	0.402	0.055	2.932
Age group (Adolescent vs Adult)	0.776	0.239	2.520
Follow-up CD4 <200 vs >200 At Treatment (Regimen1) =(NVP+D4T+3TC)	8.381	1.805	38.906
Follow-up CD4 <200 vs >200 At Treatment (Regimen1) =(EFV+D4T+3TC)	4.092	1.834	9.128

Comment on wide 95% confidence intervals for HR in Table 3.2.16:

Treatments (Regimen 1) (EFV+AZT+3TC) and (EFV+3TC+TDF) are characterised by comparatively heavy censoring (low death rates) (Table 1.6.4) and comparatively small sample sizes (Table 3.2.2), and this could explain the relatively wide confidence intervals (CIs) in univariate analysis in Table 3.2.16 (Fay et al, 2013). A similar explanation applies to other covariates with relatively wide CIs. According to (Kleinbaum and Klein, 2012), wide confidence interval indicates that the point estimate is somewhat unreliable.

3.3 Quantile regression modelling

The theory of quantile regression in survival analysis was briefly reviewed in (section 2.5) of chapter 2. The question to be answered in this section using quantile regression modelling of survival times (follow-up times) is: What are the factors that affect patients with short, medium and long survival times? It should be noted that here short, medium and long survival times respectively means survival times at quantile levels less than or equal to 0.1 ($\leq$ 9 months), above 0.1 but less than or equal to 0.5 (>9 months, $\leq$ 3.75 years) and greater than 0.5 ($>$ 3.75 years) but less than or equal to 0.9 ($\leq$ 6.75 years). The quantile regression modelling of the data to answer the question was done using Proc Quantreg and Proc Quantlife procedures in SAS Version 9.4. The Quantlife procedure performs quantile regression analysis for censored survival data while the Quantireg procedure uses quantile regression to model the effects of covariates on the conditional quantiles of a response variable. The Proc Quantreg was preferred over Proc Quantlife because its regression procedure was less restrictive, and its quantile plots were more detailed. As in theory of quantile regression in survival, quantile regression modelling in this study assumes that the conditional quantile of survival time is a linear function of covariates.

3.3.1 Quantile regression modelling at quantile levels 0.1, 0.5 and 0.9

Tables 3.3.1, 3.3.2 and 3.3.3 at the end of this section contain the results from quantile regression modelling the data. The covariates in the models were selected using an automated 'Proc QUANTSELECT' procedure in SAS version 9.4. In Table 3.3.1, the significant factors with negative significant effects on the 0.1^{st} quantile of the log (survival time) at the 0.1 level of significance are: poor and fair ART adherence relative to good ART adherence, *ln* (Follow-up viral load) and Follow-up white blood cell count while Follow-up CD4 and the interaction

effect of Follow-up white blood cell count by Treatment (Regimen 1) have positive effects. Significant factors with negative significant effects on the 0.5^{th} quantile of the log (survival time) at the 0.1 level of significance as shown in Table 3.3.2 are: WHO stage 3 relative to WHO stage 4, poor and fair ART adherence relative to good ART adherence and Baseline haemoglobin while the males relative females, Carolina hospital relative to Embhuleni hospital, WHO stage 1 relative to WHO stage 4 and Follow-up CD4 have positive effects. Lastly, Table 3.3.3 shows that significant factors with negative significant effects on the 0.9^{th} quantile of the log (survival time) at the 0.1 level of significance are: poor and fair ART adherence relative to good ART adherence and ln(Baseline viral load) while factors with positive effect are Carolina hospital relative to Embhuleni hospital, Follow-up CD4, Baseline white blood cell count and Follow-up sodium. The significant covariates like Baseline haemoglobin in quantile level 0.9 were left out from the list of significant covariates listed above. The effects at 0.1 level of significance of the left-out covariates were not significant because their 95% confidence limits contain zero which is a value of zero effect.

Table 3.3.1: Results of fitting the quantile regression model at quantile level 0.1

Parameter	Estimate	Standard Error	95% Confidence Limits		t-Value	Pr > |t|
Intercept	4072.528	937.5368	2228.2663	5916.7893	4.34	<.0001
Gender (Male vs Ref level Female)	38.1982	54.2694	-68.5569	144.9534	0.70	0.4820
Hospital (Carolina vs Ref level Embhuleni)	45.1383	87.5172	-127.0199	217.2965	0.52	0.6064
WHO Stage 1 vs Ref level WHO stage 4	225.6104	320.4300	-404.7187	855.9394	0.70	0.4819
WHO Stage 2 vs Ref level WHO stage 4	-195.365	119.5776	-430.5905	39.8602	-1.63	0.1033
WHO Stage 3 vs Ref level WHO stage 4	-83.4858	109.1579	-298.2141	131.2425	-0.76	0.4449
Treatment (Regimen 1)=1(NVP+D4T+3TC)	-1879.62	479.7574	-2823.373	-935.8772	-3.92	0.0001
Treatment (Regimen 1)=5(EFV+D4T+3TC)	-1780.55	470.7147	-2706.509	-854.5906	-3.78	0.0002
Treatment (Regimen 1)=6(EFV+AZT+3TC)	-1944.98	981.5575	-3875.832	-14.1200	-1.98	0.0484
Treatment (Regimen 1)=18(EFV+3TC+TDF)	-2013.98	565.5769	-3126.547	-901.4148	-3.56	0.0004
ART adherence (poor)	-576.204	114.1237	-800.7007	-351.7071	-5.05	<.0001
ART adherence (fair)	-470.106	104.3213	-675.3205	-264.8924	-4.51	<.0001
Follow-up CD4	0.3536	0.1192	0.1191	0.5881	2.97	0.0032
Baseline haemoglobin	-21.6111	11.3817	-44.0005	0.7783	-1.90	0.0585
Baseline lymphocyte	-1.8801	3.2958	-8.3635	4.6032	-0.57	0.5688
Baseline white blood cell count	13.7185	8.2717	-2.5530	29.9900	1.66	0.0982
Follow-up white blood cell count	-183.100	72.1291	-324.9878	-41.2125	-2.54	0.0116
ln (Baseline viral load)	-9.5667	8.5946	-26.4736	7.3401	-1.11	0.2665
Baseline sodium	-4.2570	2.7467	-9.6601	1.1461	-1.55	0.1221
Follow-up sodium	-2.3377	6.4345	-14.9952	10.3198	-0.36	0.7166
ln (Follow-up viral load)	-25.8477	9.4026	-44.3439	-7.3515	-2.75	0.0063
Follow-up WBC*(NVP+D4T+3TC)	162.4739	71.9978	20.8445	304.1034	2.26	0.0247
Follow-up WBC*(EFV+D4T+3TC)	156.1957	70.5368	17.4403	294.9511	2.21	0.0275
Follow-up WBC*(EFV+AZT+3TC)	160.5983	152.2742	-138.9457	460.1423	1.05	0.2923
Follow-up WBC*(EFV+3TC+TDF)	254.0311	77.8733	100.8438	407.2184	3.26	0.0012

Key: WBC=White Blood Cell count; **Reference levels**: (19=NVP+3TC+TDF) for Treatment (Regimen 1); 'Good' for ART adherence, Follow-up WBC*(NVP+3TC+TDF) for Follow-up WBC by Treatment (Regimen 1)

Table 3.3.2: Results of fitting the quantile regression model at quantile level

Parameter	Estimate	Standard Error	95% Confidence Limits		t-Value	Pr > |t|
Intercept	3583.654	1271.105	1083.2188	6084.0892	2.82	0.0051
Gender (Male vs Ref level Female)	102.5502	46.9079	10.2761	194.8244	2.19	0.0295
Hospital (Carolina vs Ref level Embhuleni)	346.9829	114.6823	121.3873	572.5785	3.03	0.0027
WHO Stage 1vs Ref level WHO stage 4	575.5814	239.3157	104.8151	1046.3478	2.41	0.0167
WHO Stage 2vs Ref level WHO stage 4	-215.944	111.2944	-434.8751	2.9873	-1.94	0.0532
WHO Stage 3vs Ref level WHO stage 4	-195.391	98.8015	-389.7470	-1.0353	-1.98	0.0488
Treatment (Regimen 1)=1(NVP+D4T+3TC)	-1583.98	570.6746	-2706.573	-461.3846	-2.78	0.0058
Treatment (Regimen 1)=5(EFV+D4T+3TC)	-1440.93	590.1048	-2601.741	-280.1094	-2.44	0.0151
Treatment(Regimen 1)=6(EFV+AZT+3TC)	-464.181	1049.961	-2529.596	1601.2333	-0.44	0.6587
Treatment (Regimen 1)=18(EFV+3TC+TDF)	-520.450	732.9321	-1962.226	921.3266	-0.71	0.4781
ART adherence (poor)	-909.383	114.0102	-1133.656	-685.1092	-7.98	<.0001
ART adherence (fair)	-771.606	115.3228	-998.4619	-544.7508	-6.69	<.0001
Follow-up CD4	0.5993	0.1768	0.2516	0.9470	3.39	0.0008
Baseline haemoglobin	-28.3134	10.5987	-49.1626	-7.4643	-2.67	0.0079
Baseline lymphocyte	-5.4162	3.4098	-12.1237	1.2914	-1.59	0.1131
Baseline white blood cell count	15.2359	10.5539	-5.5250	35.9968	1.44	0.1498
Follow-up white blood cell count	-52.8942	101.9549	-253.4533	147.6648	-0.52	0.6042
ln (Baseline viral load)	-18.8115	11.0841	-40.6154	2.9923	-1.70	0.0906
Baseline sodium	-3.5823	3.2782	-10.0310	2.8664	-1.09	0.2753
Follow-up sodium	2.0567	8.4248	-14.5160	18.6293	0.24	0.8073
ln (Follow-up viral load)	-4.8583	10.7122	-25.9306	16.2141	-0.45	0.6505
Follow-up WBC*(NVP+D4T+3TC)	40.8608	98.5437	-152.9881	234.7096	0.41	0.6787
Follow-up WBC*(EFV+D4T+3TC)	20.2545	101.8067	-180.0131	220.5221	0.20	0.8424
Follow-up WBC*(EFV+AZT+3TC)	-87.3588	154.1302	-390.5538	215.8362	-0.57	0.5712
Follow-up WBC*(EFV+3TC+TDF)	-1.6828	136.3815	-269.9636	266.5979	-0.01	0.9902

Key: WBC=White Blood Cell count; **Reference levels**: (19=NVP+3TC+TDF) for Treatment (Regimen 1); 'Good' for ART adherence, Follow-up WBC*(NVP+3TC+TDF) for Follow-up WBC by Treatment (Regimen 1)

Table 3.3.3: Results of fitting the quantile regression model at quantile level 0.9

Parameter	Estimate	Standard Error	95% Confidence Limits		t-Value	Pr > |t|
Intercept	-177.192	2846.160	-5775.972	5421.5886	-0.06	0.9504
Gender (Male vs Ref level Female)	154.0139	137.0287	-115.5402	423.5680	1.12	0.2618
Hospital (Carolina vs Ref level Embhuleni)	733.1584	172.4150	393.9948	1072.3219	4.25	<.0001
WHO Stage 1 vs Ref level WHO stage 4	4.3415	330.9289	-646.6403	655.3233	0.01	0.9895
WHO Stage 2 vs Ref level WHO stage 4	-565.918	346.5517	-1247.633	115.7955	-1.63	0.1034
WHO Stage 3 vs Ref level WHO stage 4	-461.651	238.8814	-931.5633	8.2605	-1.93	0.0541
Treatment (Regimen 1)=1(NVP+D4T+3TC)	-805.346	1264.085	-3291.973	1681.2804	-0.64	0.5245
Treatment (Regimen 1)=5(EFV+D4T+3TC)	-629.681	1258.636	-3105.589	1846.2264	-0.50	0.6172
Treatment (Regimen 1)=6(EFV+AZT+3TC)	768.8774	1821.161	-2813.592	4351.3467	0.42	0.6732
Treatment (Regimen 1)=18(EFV+3TC+TDF)	746.0460	1448.810	-2103.959	3596.0510	0.51	0.6069
ART adherence (poor)	-954.615	184.5548	-1317.660	-591.5712	-5.17	<.0001
ART adherence (fair)	-392.288	164.6356	-716.1483	-68.4274	-2.38	0.0177
Follow-up CD4	0.9745	0.3796	0.2278	1.7212	2.57	0.0107
Baseline haemoglobin	-44.1973	23.9782	-91.3656	2.9709	-1.84	0.0662
Baseline lymphocyte	-9.6539	7.9296	-25.2525	5.9447	-1.22	0.2243
Baseline white blood cell count	44.3816	18.5188	7.9526	80.8106	2.40	0.0171
Follow-up white blood cell count	-37.0213	227.0659	-483.6907	409.6480	-0.16	0.8706
ln (Baseline viral load)	-44.9037	19.3369	-82.9419	-6.8655	-2.32	0.0208
Baseline sodium	-10.7307	7.4767	-25.4384	3.9769	-1.44	0.1522
Follow-up sodium	39.9021	16.6911	7.0684	72.7358	2.39	0.0174
ln (Followup viral load)	-6.9361	21.2915	-48.8194	34.9472	-0.33	0.7448
Follow-up WBC*(NVP+D4T+3TC)	-42.1296	228.7746	-492.1601	407.9010	-0.18	0.8540
Follow-up WBC*(EFV+D4T+3TC)	-37.2601	226.4757	-482.7684	408.2483	-0.16	0.8694
Follow-up WBC*(EFV+AZT+3TC)	-152.768	290.1790	-723.5889	418.0537	-0.53	0.5989
Follow-up WBC*(EFV+3TC+TDF)	-160.294	257.6449	-667.1161	346.5284	-0.62	0.5343

Key: WBC=White Blood Cell count; **Reference levels**: (19=NVP+3TC+TDF) for Treatment (Regimen 1); 'Good' for ART adherence, Follow-up WBC*(NVP+3TC+TDF) for Follow-up WBC by Treatment (Regimen 1)

Table 3.3.4: Likelihood Ratio and Wald Tests of significance of the quantile models in the tables 3.3.1, 3.3.2 and 3.3.3

Test Results					
Quantile Level	Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
0.1	Wald	98.4996	20	98.50	<.0001
0.1	Likelihood Ratio	188.4828	20	188.48	<.0001
0.5	Wald	1140.2448	20	1140.24	<.0001
0.5	Likelihood Ratio	567.1370	20	567.14	<.0001
0.9	Wald	333.3253	20	333.33	<.0001
0.9	Likelihood Ratio	279.2965	20	279.30	<.0001

Table 3.3.4 shows that both Wald test and likelihood ratio test indicate that the coefficients of the models are significantly different from zero at the 10^{th}, 50^{th} and 90^{th} quantiles.

3.3.2 Some quantile process plots to study heterogeneity in the data

A visual comparison of the magnitudes of the effects of each covariate in Figures 3.3.1 and 3.3.2 (or quantile levels 0.1, 0.5, 0.9) suggests the existence of heterogeneity in the data. In this section, this is formally investigated using quantile process plots for (effects versus quantile levels) Follow-up CD4, *ln* (Baseline viral load) and Treatment (Regimen 1). These covariates were significant either as main effects or as interaction effects in all of Cox regression, logistic regression and quantile regression modelling while ART adherence was significant in both Cox regression and quantile regression modelling. The quantile process plots are displayed in Figures 3.3.1 to 3.3.3.

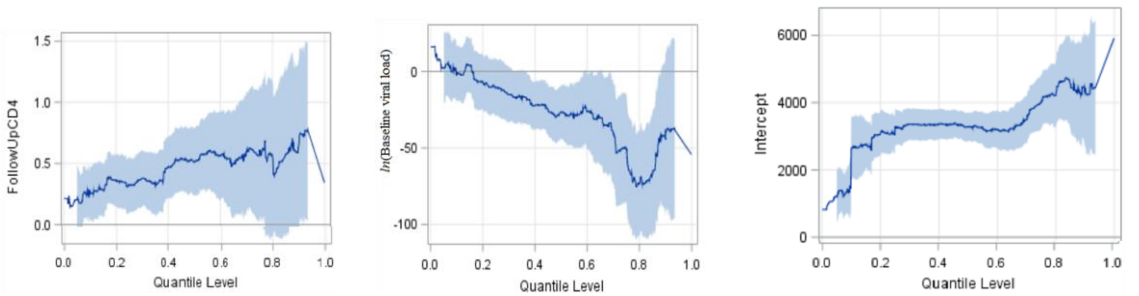


Figure 3.3.1: Estimated parameter versus quantile for log (survival time) with 95% confidence limits - Follow-up CD4, *ln (*Baseline viral load) and intercept

Figure 3.3.1 shows that Follow-up CD4 has a positive effect on log (survival time) which increases with increase in the quantile levels. *ln* (Baseline viral load) has a negative effect on log (survival time) which increases in magnitude with increase in the quantile levels up to the 0.8th quantile. Then the effect diminishes in magnitude. Thus, the effects of Follow-up CD4 and *ln (*Baseline viral load) on log (survival time) are quantile specific. This is also true for the intercept whose positive effect on log (survival time) increases with increase in the quantile levels. A closer look at the relationship between log (survival time) and *ln (*Baseline viral load) (expressed by the slope) shows lack of significance for quantiles below 0.35th since zero lies within the 95% confidence limits.

Finally, Figure 3.3.2 displays quantile process plots for ART adherence and Treatment (Regimen 1).

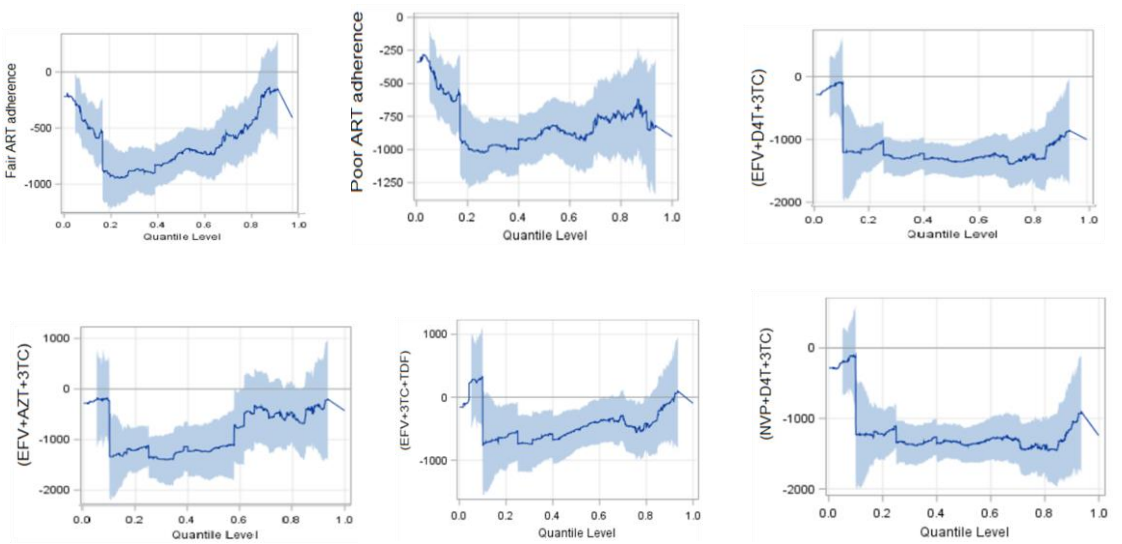


Figure 3.3.2: Estimated parameter by quantile for *ln* (survival time) with 95% confidence limits - ART adherence and Treatment (Regimen 1)

Figure 3.3.2 shows that (EFV+3TC+TDF), (EFV+D4T+3TC), (EFV+AZT+3TC), (NVP+D4T+3TC), Poor ART Adherence, and Fair ART Adherence each has a sudden, negative and worsening effect on log (survival time) up to around 0.3rd quantile level. As from around 0.3rd quantile upwards, the negative effect diminishes in magnitude for (EFV+3TC+TDF), (EFV+AZT+3TC), poor ART adherence and fair ART adherence while the diminishing effect for (EFV+D4T+3TC) and (NVP+D4T+3TC) takes place as from around the 0.8th quantile upwards. Thus, the effects of ART adherence and Treatment (Regimen 1) on *ln* (survival time) are quantile specific.

3.3.3 Survival analysis of the effect of covariates on quantiles of the survival time of patients

According to Lin and Rodriguez (2013), the applications of quantile regression to survival analysis include studying the effect of a specific covariate on the survival time of an individual. Quantile regression in (subsection 3.3.1) has given us a picture of covariates which affect most patients with short, median and long follow-up time. The quantile regression analysis presented herein as in (section 3.3.2), are for Follow-up CD4, *ln* (Baseline viral load), Treatment (Regimen 1) and ART adherence. A given covariate may have a different effect on the survival of patients with low, medium and long follow-up time as shown in Figure 3.3.3 and Figure 3.3.4. Figure 3.3.3 shows that an increase in Follow-up CD4 count results in high improvement in survival of patients in the 0.5^{th} quantile while a gradual improvement is seen in 0.1^{st} and 0.9^{th} quantiles.

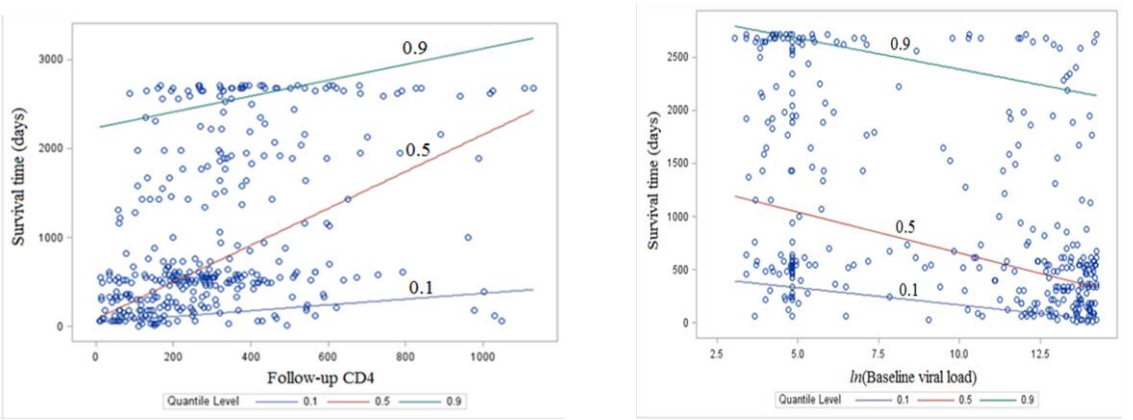


Figure 3.3.3: Survival analysis of a continuous covariate effect on quantiles of Survival time

In terms of survival, HIV/AIDS patients on ART for around 3.75 years respond better to increase in Follow-up CD4 count than patients on ART for any other duration within the 7.5 years. An increase in Baseline viral load is accompanied by a decrease in survival for all the quantiles of Survival time. In terms of survival, HIV/AIDS patients on ART for around 3.75 years are the most negatively affected by an increase in Baseline viral load than patients on ART for any other duration within the 7.5 years.

Figure 3.3.4 shows the effects of categorical covariates on quantiles of survival time. In terms of survival, HIV/AIDS patients on ART for around 3.75 years respond better to Treatment (Regimen 1) than patients on ART for any other duration within 7.5 years. Treatments

(EFV+3TC+TDF) and (NVP+3TC+TDF) are associated with higher survival at all quantiles than Treatments (NVP+D4T+3TC), (EFV+D4T+3TC) and (EFV+AZT+3TC). HIV/AIDS patients on ART for around 3.75 years respond better to fair or good ART adherence than patients on ART for any other duration within 7.5 years.

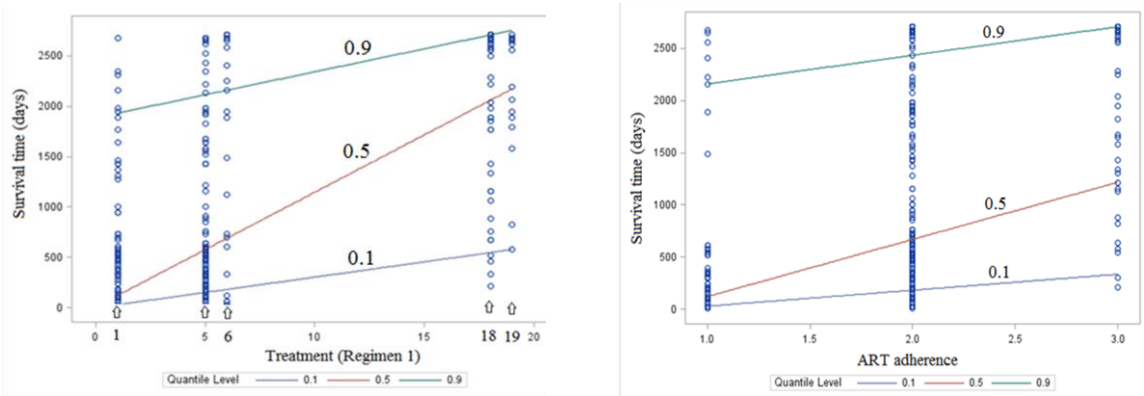


Figure 3.3.4: Survival analysis of a categorical covariate effect on quantiles of Survival time

Key: **Treatment (Regimen 1):** 1 = (NVP+D4T+3TC), 5 = (EFV+D4T+3TC), 6 =(EFV+AZT+3TC), 18 = (EFV+3TC+TDF) and 19 = (NVP+3TC+TDF).
ART adherence: 1 = poor ART adherence, 2 = Fair ART adherence and 3 = Good ART adherence

Good ART adherence is associated with better survival than fair ART adherence which in turn is associated with better survival than poor ART adherence at all quantiles.

3.4 Summary

Chapter three presented the results of the study in view of the research questions and objectives. The chapter focused on identification, modelling and discussions of the factors affecting the survival lifetime of HIV+ terminal patients in Albert Luthuli Municipality of South Africa using Logistic regression, Cox regression and Quantile regression approaches. Tables, figures, and short descriptions were used as tools for data analysis.

CHAPTER 4

FINDINGS AND DISCUSSIONS

4.1 Findings

Chapter four presents research findings and discussions based mainly on data analysis in Chapter three, and in comparison to some previous related studies. The findings in this study could be useful for health practitioners in the treatment of HIV/AIDS patients, in counselling patients about their prognosis and in policy formulation regarding HIV/AIDS management.

4.1.1 Findings based on demographic and outcome factors

The status of all the 357 patients at the end of the follow-up as in Table 3.1.1, shows that about half of the total number of patients (49.3%) transferred to other health care institutions outside Albert Luthuli Municipality, about a quarter of patients were lost to follow-up and slightly above one in ten patients were alive and still attached to a hospital at the end of the study. The percentage of HIV/AIDS patients lost to follow-up from the sample of 357 was about 26% (mostly from adolescents) and this is about five-fold the result from similar studies done by Sieleunou et al. (2008) in Far-North Province of Cameroon from 2001 to 2006. In similar studies, Damtew et al. (2015) and Moshago et al. (2014) also concluded that a higher proportion of adolescents was lost to follow-up in comparison to child and adult age groups. The overall mortality among the 357 HIV/AIDS patients in this study was around 11.5% (mostly from adults), and this nearly concurs with the similar study findings by Tadege (2018) and Damtew et al. (2015) wherein around 15% and 14.5 % patients died respectively. However, the death rate in this study is double the mortality in Tshwane district as per similar study carried out by Mlangeni and Senkubuge (2016). Table 3.1.1 also shows that most HIV/AIDS patients got their treatment from Embhuleni hospital and the mortality at Embhuleni hospital was slightly more than twice the mortality at Carolina hospital. The proportions of HIV/AIDS patients getting treatment in Albert Luthuli Municipality shows that females constitute a greater percentage as shown in Table 3.1.2. The proportion of males lost to follow-up was greater than that for females, while the proportion of females who transferred from the hospitals was greater than that for males as shown in Table 3.1.2. Table 3.1.3 also confirms that children had the highest transfer percentage out of the hospital (60%) and they also had the greatest percentage

of being still alive (20%) at the end of study. In addition, Table 3.1.3 shows that Adults constituted more than 80 % of HIV/AIDS patients who visited the two hospitals for treatment. The long-term retention of HIV/AIDS patients in Albert Luthuli was around 13 % for the 2010-2017 period as shown in Table 3.1.3, and this sharply contrasts to 57% which was the long-term retention in Tshwane district for the 2007-2011 period.

4.1.2 Findings based on Logistic regression

The findings related to Logistic regression present the following as factors which are associated with mortality among the HIV+ terminal patients: Follow-up CD4, Marital status, Follow-up lymphocyte, Baseline viral load, Follow-up Alanine Transaminase and the interaction of Follow-up mass by Treatment (Regimen 1), as shown in Table 3.1.4. Among these factors, Follow-up CD4 and Baseline viral load are the most highly significant factors ($p - value < 0.005$). Another finding of interest is that, the researched independent factors which are associated with Lost to follow-up and Transfer from hospital as shown in Table 3.1.5 are: Treatment (Regimen 1), ART adherence, Follow-up haemoglobin, Baseline lymphocyte and the interaction of Age by Baseline viral load. Treatment (Regimen 1) and ART adherence are the most highly significant factors ($p - value < 0.005$) associated with the transfer and loss of patients to follow-up.

4.1.3 Findings based on Cox regression

The Cox regression as the main tool for data analysis in this study, identified as shown in Table 3.2.1, the factors affecting the survival lifetime of HIV+ terminal patients as: ART adherence, Age, Follow-up mass, Baseline sodium, Baseline viral load and interactions of Follow-up lymphocyte by TB history and Follow-up CD4 by Treatment (Regimen 1). The two most highly significant factors ($p - value < 0.001$) in this finding were ART adherence and Baseline viral load. Poor ART adherence, Baseline sodium, advanced Age and Baseline viral load are associated with poor survival experience, while an increase in Follow-up mass and good ART adherence are associated with improved survival experience. The interaction Follow-up CD4 by Treatment (Regimen 1) is associated with improved survival experience, while the interaction Follow-up lymphocyte by TB history is associated with poor survival experience among the HIV+ terminal patients. In similar studies, Damtew (2015) had WHO stage, CD4 cell count, haemoglobin, weight and concomitant TB infection as the factors

associated with survival experience, while Teka et al. (2015), had factors associated with survival as age, formal education, family size, alcoholic consumption, tobacco and chat use, baseline weight, current weight, baseline CD4 cell count, baseline haemoglobin, and TB status. The covariates from these similar studies which concur with the findings of this study are: CD4 cell count, weight, concomitant TB infection and age. The results from similar studies by Shebeshi (2011) had age, CD4 count, WHO stage and Adherence as significant covariates, while from Moshego (2012) the significant covariates were CD4 count, TB infection and WHO stage. Lastly, in similar studies in Johannesburg, Sanne et al. (2009), had predictors of survival which concur with the findings of this study as ART regimens, CD4 count and viral load.

4.1.4 Findings based Kaplan-Meier survival functions and Hazard ratios

The findings herein are based on Kaplan-Meier and hazard ratios for the factors which have been found to be statistically significant and to have significant hazard ratios from Cox regression. The following findings are derived from Table 3.1.16 and in relationship to Kaplan-Meier survival function estimates in Subsection 3.2.2 of chapter 3. By keeping all other factors constant in each case, the following baseline findings hold for the study of cohort of HIV+ terminal patients.

4.1.4.1 ART adherence

In this study the ART adherence levels which are: poor, fair and good match the WHO guideline ART adherence levels which are: low adherence, moderate adherence and high adherence respectively. ART adherence in this study was considered in the same way Shebeshi (2011) viewed it; that is taking all ARV pills in the correctly prescribed doses at the right time and in the right way observing any dietary restriction. As shown in Table 1.6.4, the percentages of patients who died, while on poor adherence, fair adherence or good adherence were 48.8%, 46.3% and 4.9% respectively. The summary in Table 1.6.4 shows us that most deaths are associated with poor ART adherence. As reflected in Figure 3.2.5 and Table 3.2.16, HIV/AIDS patients with poor ART adherence relative to patients with fair ART adherence are about 6 times likely to experience death as a result of HIV related illness, while patients with poor ART adherence in comparison to patients with good ART adherence are about 18 times likely to experience the same event.

4.1.4.2 Age

Adolescent HIV/AIDS patients experienced the lowest survival probability (as low as around 82%) from around the 1000th to around the 2200th day of follow-up. This lowest survival probability could prompt a higher loss to follow-up as indicated in section 4.1.1. Adult patients experienced lowest survival probability (around 82%) from around the 2200th to around the 2500th of the survival time. Child patients experienced the highest survival probability (100%) throughout the follow-up period except for the period from the 2500th day to final follow-up day when they experienced the lowest survival probability (around 80%). Could this high survival among children be as a result of high transfer rate as indicated in Section 4.1.1? As shown in Figure 3.2.6, adult and adolescent patients experienced lower survival probability than child patients. Despite the differences in these age groups as described above, Log rank Table 3.2.9 shows that Age strata in Figure 3.2.6 are not statistically different (p-value<0.6010) and this is confirmed by the crossing of the respective graphs and by non-significant hazard ratios in Table 3.2.16. This passage serves the purpose that real life differences in phenomenon may not necessarily have statistical difference or the other way around.

4.1.4.3 Follow-up mass

As shown in Figure 3.2.7, HIV/AIDS patients with normal Follow-up mass have the highest survival probabilities, while patients with Follow-up mass below normal have the lowest survival probabilities. HIV/AIDS patients with Follow-up mass below normal in comparison to patients with follow-up mass above normal are about twice likely to experience death as a result of HIV related hazard.

4.1.4.4 Baseline viral load

The main aim of HIV treatments is to reduce the viral load to a point where the viral load is undetectable in the blood, and it is no longer attacking the immune system (Fletcher, 2018). This gives the immune system the chance to rebuild its healthy cell count and results in increase in the number of CD4 cells as viral load falls, as explained in Appendix H. Figure 3.2.9 shows that HIV/AIDS patients with Baseline viral load above 5000 HIV RNA copies/mm^3 have the lowest survival probability, while patients with Baseline viral load below 50 HIV RNA copies/ mm^3 (undetectable viral load) have the highest survival probability (100%) throughout the

follow-up period. Hazard ratios in Table 3.2.16 show us that patients with Baseline viral load between 50 HIV RNA copies/mm^3 and 5000 HIV RNA copies/mm^3 have HIV hazard which around 87% lower relative to patients with Baseline viral load above 5000 HIV RNA copies/ mm^3.

4.1.4.5 Follow-up lymphocyte by TB history groups

In this study, the risk of death for patients with tuberculosis (TB) (36.6% of died patients in Table 1.6.4) is about 3.5 times higher relative to patients who were TB negative (Table 3.2.16). This finding is close to similar study by Tadege (2018) where the risk of death for patients who lived with tuberculosis was about 2.9 times higher relative to patients who were TB negative. Patients with TB history in comparison to patients without TB history at Follow-up lymphocyte below normal are about 3 times likely to experience death as a result of HIV related illness. Keeping all other covariates constant, patients with low Follow-up lymphocyte and with concomitant TB have the lowest survival experience, and they experienced the event (death) earlier than other groups. In addition, patients belonging to this stratum died before the end of the study. Patients with normal Follow-up lymphocyte level and without TB history have the highest survival experience if other covariates are kept constant. As shown in Table 3.2.16, Patients with TB relative to those without (at Follow-up lymphocyte below normal) are about 2.8 times likely to experience HIV related hazard (death). According to Moshego et al. (2012), similar studies on retrospectively followed HIV/AIDS patients on ART demonstrated that HIV-infected patients with TB positive had shorter survival duration.

4.1.4.6 Follow-up CD4 by Treatment (Regimen 1)

According to Moshego et al. (2012), the results from different researchers showed that CD4 cell counts had a strong influence on the survival experience of HIV/AIDS patients on ART. As shown in Figure 3.2.12, HIV/AIDS patients with Follow-up CD4 count below 200 cells/mm^3 experienced lower survival probability than patients with Follow-up CD4 count above 500 cells/mm^3. Table 3.2.16 shows that HIV/AIDS patients with Follow-up CD4 count below 200 cells/mm^3 relative to patients with Follow-up CD4 count between 200 cells/mm^3 and 500 cells/mm^3 are about 5.5 times likely to experience death as a result of HIV related hazards, while patients with Follow-up CD4 count below 200 cells/mm^3 relative to patients with Follow-up CD4 count above 500 cells/mm^3 are about 26 times likely to experience death.

The condition of HIV disease when CD4 is below 200 cells/mm^3 is AIDS. Figure 3.2.13 shows that, on ART initiation, AIDS cases were 254 (71.1%) and this dropped to 139 (38.9%) at the end of the follow-up. This is a 32.2% decrease in AIDS cases over 7.5 years as a result of the intervention effect of ART administration on CD4 immunology, which is supposed to have an inverse relationship with virology (viral load) as explained in Appendix G.

As shown in Table 1.6.4, the percentages of patients who died using (NVP+D4T+3TC), (EFV+D4T+3TC), (EFV+AZT+3TC), (EFV+3TC+TDF) or (NVP+3TC+TDF) as their first-line regimen were 26.8%, 65.9%, 4.9%, 2.4% and 0% respectively. The TDF-containing regimens are associated with low death rate. Treatment regimen (NVP+D4T+3TC) relative to treatment regimen (EFV+3TC+TDF) is about 14 times, while treatment regimen (EFV+D4T+3TC) relative to treatment regimen (EFV+3TC+TDF) is about 15 times likely to give rise to death as a result of HIV related hazards. Patients on (NVP+D4T+3TC) treatment relative to patients on (EFV+3TC+TDF) treatment at Follow-up CD4 count below 200 are about 14 times likely to experience death as a result of HIV related hazards, while patients on (EFV+D4T+3TC) treatment relative to patients on (EFV+3TC+TDF) treatment at Follow-up CD4 count below 200 are about 15.5 times likely to experience death. Patients with Follow-up CD4 above 200 cells/mm^3 and taking (EFV+3TC+TDF) or (EFV+D4T+3TC) have the highest survival experience, while on the other hand patients with Follow-up CD4 below 200 cells/mm^3 and taking (EFV+AZT+3TC) have the lowest survival experience as shown in Figure 3.2.11. According to Ochieng-Ooko et al. (2010), ART (Treatment) has significantly reduced mortality and improved life expectancy.

4.1.4.7 Follow-up sodium

Hyponatremia disease is a low sodium concentration (less than 135 mmol/L) in the blood and it is a marker of severity of HIV-disease but not an independent risk factor for mortality (Braconnier et al., 2017). As shown in Figure 3.2.8, HIV/AIDS patients with Baseline sodium below normal have the lowest survival probability. However, the pairwise comparisons of the Baseline sodium strata as shown in Table 3.2.11 and Table 3.2.16 show no statistical difference in the survival experiences of HIV/AIDS patients.

4.1.5 Findings based on Quantile regression

The HIV/AIDS patients with short follow-up times (<0.1 level or ≤ 9months) as shown in Table 3.3.1 are affected most by: the interaction of Follow-up white blood cell count by regimen 1 treatments (EFV+3TC+TDF), (NVP+D4T+3TC) or (EFV+D4T+3TC) relative to Treatment (NVP+3TC+TDF), Follow-up CD4, poor and fair adherence relative to good adherence and *ln* (Follow-up viral load).

The HIV/AIDS patients with median follow-up times (<0.5 level or ≤3.75 years) are affected most by: Gender, Hospital, poor and fair adherence relative to good adherence, Follow-up CD4, Baseline haemoglobin, and WHO Stages 1, and 3 relative to WHO stage 4.

Lastly, patients with long follow-up times (<0.9 level or ≤6.75 years) are affected most by: Hospital, poor and fair ART adherence relative to good adherence, Follow-up CD4, *ln* (Baseline viral load), Follow-up sodium and Baseline white blood cell count.

ART adherence and Follow-up CD4 are significant at all the three quantile levels, while Hospital is significant at Quantile levels 0.5 and 0.9. Thus, these three covariates are not as sensitive to quantile levels as some other significant variables used in this study.

A formal investigation of Follow-up CD4, *ln* (Baseline viral load), Treatment (Regimen 1) and ART adherence among other factors, by using quantile process plots for (effects versus quantile levels) suggests the existence of heterogeneity in the data. Survival analysis of the effects of individual covariates on quantiles of the survival time of patients by using fit plots reflected that an increase in Follow-up CD4 count is associated with an improvement in survival experience, while an increase in Baseline viral load is associated with poor survival experience among the HIV/AIDS patients. On the other hand, the quantile fit plots on Treatment (Regimen 1) depicted that TDF-containing regimens are associated with better survival experience than EFV-containing regimens. The EFV-containing regimens in turn are associated with better survival experience than NVP-containing regimens. The fit plots also depicted that patients with good ART adherence experience are associated with better survival experience than patients with fair ART adherence who in turn are associated with better survival experience than patients with poor ART adherence.

4.1.6 Findings based on SAS output for 95% hazard ratio confidence intervals for small samples and heavy censoring

The covariates which are characterised by small sample sizes and/or heavy censoring (low death rate) in this study have been found to have extremely wide hazard ratio confidence intervals. This is the case with Treatments (Regimen 1) (EFV+AZT+3TC) and (EFV+3TC+TDF) in Table 3.2.16 and (NVP+3TC+TDF) in Table 3.2.7.

4.2 Discussions

The findings in this study might not adequately compare with some related studies because of the following reasons:

i) In the case of Damtew et al. (2015) and Moshago et al. (2014), the focus was mainly on death as the survival threat. This study, in addition to considering death as a threat to HIV/AIDS patients, it also considered patient transfer out of hospital and loss of patients to follow-up as phenomena which affect the long-term success of antiretroviral therapy programme and ultimately leading to increased death cases among HIV/AIDS patients especially in the case of lost to follow-up.

ii) The national strategic plan of 2012-2016 by the Department of Health of South Africa stipulates that the retention rate of patients on ART should be at least 80% (Mlangeni & Senkubuge, 2016). The long-term retention of HIV/AIDS patients in Albert Luthuli was around 13% for the 2010-2017 period; this sharply deviates from the national strategic plan of 2012-2016. This sharp deviation in retention rates could be explained by challenges experienced in Albert Luthuli in form of high patient transfers (at 49.3%) and high patient loss to follow-up (at 26%).

iii) Related studies done by Damtew et al. (2015) and Sieleunou et al. (2008) focused on adult patients and on patients with ages greater than 15 years. This study is related to the study by Moshago et al. (2014) in that the sample was exhaustively handled by dividing it into children, adolescents and adults. This was done with the view to find a solution to all groups at once using the limited resources available.

iv) Most similar studies which were carried out were limited mostly to a follow-up period of 4-5 years as in the case of Sieleunou et al. (2008), Damtew et al. (2015) and Mlangeni

& Senkubuge (2016). This research study spanned over a longer follow-up period of 7.5 years with the view to improve on the quality of the findings.

v) The sample size for this research closely matches the sample sizes for studies by Tadege (2018), Teka et al. (2015), and Mlangeni & Senkubuge (2016), however, it is comparatively smaller than the sample sizes for studies by Shebeshi (2011), Damtew et al. (2015) and Sieleunou et al. (2008) who retrospectively followed greater numbers of HIV/AIDS patients which were 456; 784 and 1187 respectively.

vi) This study reports a comparatively high percentage in terms of patients lost to follow-up and this sharply contrasts to the report by Sieleunou et al. (2008) on HIV/AIDS patients in Cameroon where some intensive follow-up to those lost to follow-up was carried out. The follow-up in Albert Luthuli Municipality involves phone call, whereas in Cameroon follow-up involves phone calls and messengers.

Furthermore, one of the assumptions used in survival analysis is that patients who drop out or are censored have the same survival prospects as those who continue to be followed. This assumption may theoretically hold, but from the point of view of the practicalities of this study, patients lost to follow-up were defaulters to ART adherence and hence were presumed to be at higher risk to HIV/AIDS than those who were followed continuously.

vii) Unlike most similar studies carried in South Africa for example by Mlangeni & Senkubuge (2016), and other countries which concentrated most on urban institutions; this study just like the study by Sieleunou et al. (2008), analysed the outcomes of ART in a marginalised rural set-up which presents its own peculiar challenges.

viii) This research focused on complete hospital records only as independent factors, that means any other factors with the potential to affect the survival of HIV+ terminal patients which were not part of the hospital records or were incomplete hospital records (for example records on hypertension and diabetes) were not considered in this research.

ix) Unlike in similar studies by Damtew (2015), Moshego (2012) and by Shebeshi (2011) where WHO stage emerged as a significant covariate, in this study it was not. The difference, the discrepancy or lack of statistical significance could be due to difference in calibration of the research tools, could be attributed to sample size or could probably be as a result of difference in research designs.

The application of quantile regression in identification of how covariates effects vary across patient quantiles, warrants the need for differential treatment based on lower tail (shorter length of follow-up time of a patient), median (median length of follow-up time of patients) and on upper tail (long length of follow-up time of patients). Thus, the study of quantile process plots may be applied in designing and timing of treatments which suit patients with short, medium or long follow-up times.

This study witnessed unemployment (as shown in Appendix G), rural location for the patients as shown in Appendix F and poverty (as shown by high dependency ratio in Appendix G) as factors contributing to poor ART adherence in Albert Luthuli Municipality. Poor adherence is associated with loss in follow-up of the patients as shown in Table 3.1.5. Some patients, for example farm workers, may fail to access drugs because they do not have money for transport as in a similar study by Azia et al. (2016). Farm workers may also be affected by the remoteness of the place or by the conditions of employment which may not promote frequent absenteeism. If Wellness centres in Albert Luthuli could be run even during weekends, that will enable most workers to access their ART supplies conveniently. Alternatively, or in addition, social workers could be assigned the responsibility to distribute ART drugs to remote areas like farms. Settling of these factors will go a long way in redressing poor ART adherence and lost to follow-up. Mlangeni and Senkubuge (2016) pointed out that, while the ART programme is expanding in SA, it is therefore important to monitor the programme on a long-term basis by following cohort outcomes. Croxford et al. (2016) highlighted the importance of prompt diagnosis, care engagement, and optimum management of comorbidities in reducing mortality in people with HIV.

The average mass of an adult South African is 65.7 kg (Walpole et al., 2012; “Human body weight,” 2018). Following the ART administration to the 357 participants in Albert Luthuli Municipality, 61.1 %, 6.7% and 32.2% of the patients were underweight, normal weight and overweight, respectively. In a similar study in rural community in Limpopo; Mashinya et al. (2016) found out that, of the 214 participants, 8.9%, 54.7% and 36.4% were underweight, normal weight and overweight, respectively. These two studies concur on the issue of extreme weight problems and high concurrence is on the issue of overweight. Follow-up mass was modelled as a significant covariate in both Logistic regression and Cox regression models of

this study. There is increasing evidence that obesity is an independent risk factor for the progression of Chronic Kidney Damage (CKD) (Slack et al., 2010). Mashinya et al. (2016), pointed out the need to simultaneously address the two extreme weight problems in people on ART through educating the vulnerable population on the benefits of staying away from tobacco use, engaging in physical activity and increasing awareness on cardiovascular risk. A good practice in HIV/AIDS management in Albert Luthuli Municipality district hospitals is that, a patient is initiated on ART after a free consultation of a Dietician for the provision and counselling to address the need to eat balanced diet which helps in weight challenges.

The application of Logistic regression in this study modelled Follow-up lymphocyte and Follow-up Alanine Transaminase as some of the covariates which contributed significantly to the mortality of HIV/AIDS patients. Alanine Transaminase and Lymphocyte are described in Appendix I. Low level of Lymphocyte count is independently associated with increased rate of progression of Chronic Kidney Damage (CKD) (Kim & Kim, 2014), while elevation in Alanine Transaminase is predictive of increased mortality from liver (Nagu et al., 2012). There is likelihood of HIV+ terminal patients to experience kidney dysfunctionality as a result of ART drugs. This is supported by the descriptive statistics of this study wherein 53 out 357 patients experienced kidney problems and some patients even died due to problems associated to kidney damage. Boswell and Rossouw (2017) reported manifestation of TDF nephrotoxicity within the first 3-9 months from ART initiation. Eneyew et al. (2016) in agreement with Crum-Cianflone et al. (2010) and Estrella et al. (2010) pointed out that the prevalence of renal dysfunction is projected to increase as the life expectancy of the HIV-infected individuals increases due to drug therapy.

In this study, as shown in Table 1.6.2 of chapter 1, most patients (61.9%) died within 1 year from ART initiation and this concurs with similar studies by Shebeshi (2011) in which most of the HIV/AIDS patients also died during the first 12 months. In addition, this study has found out that the greatest number of patients died within the interval from 6 months to 1 year. These findings on the distribution of mortality among HIV/AIDS patients are important in the timing of the treatment interventions.

The Kaplan-Meier and hazard ratios in this study have shown that HIV/AIDS male patients have lower survival experience than HIV/AIDS female patients. Men are far more likely to die from an HIV-related illness than women, although women are becoming infected with HIV at

a much faster rate (Merab, 2018). According to Dr Mwau in Merab (2018), the reasons why men are far more likely to die from an HIV-related illness than women are: poor health-seeking behaviour, poor social network, cultural socialisation and alcohol abuse. These same reasons are highly likely to contribute significantly to similar scenario in Albert Luthuli Municipality of South Africa. In addition, Ochieng-Ooko et al. (2010), pointed out that late treatment initiation predisposes men to a poor clinical outcome and to being lost to follow-up.

Treatments (Regimen 1) (EFV+AZT+3TC) , (EFV+3TC+TDF) and (NVP+3TC+TDF) are characterised by comparatively heavy censoring (low death rates) (Table 1.6.4) and comparatively small sample sizes (Table 3.2.2), and this could explain the relatively wide confidence intervals in univariate analysis in Table 3.2.16. and in multivariate analysis in Table 3.2.7. Small samples and/or heavy censoring could be the reason for the relatively wide CIs (Fay et al.,2013). The 95% confidence intervals (CIs) for the estimated survival function are computed using the estimate of the standard error as in equations (6) and (7). The hazard ratio confidence intervals which are computed using SAS by applying the Greenwood's formula (equations (6) and (7)), has its variables as sample size and number of deaths. In situations where the number of the observed failures is not large and/or there are few remaining subjects at risk, the CIs for Kaplan–Meier estimator can have problems (Fay et al., 2013). Small sample and/or heavy censoring could be the reason for the relatively wide CIs as opposed to possible reasons in form of outliers and collinearity which were fully addressed in this study. A similar explanation applies to other covariates with relatively wide CIs. Fay et al. (2013) propose the application of an R package called beta product confidence procedure (BPCP) in handling pointwise confidence intervals for a survival distribution with small samples or heavy censoring. BPCP is a non-parametric confidence procedure for the survival curve at a fixed time for right-censored data assuming independent censoring.

Although this study considered participants from Albert Luthuli municipality, the nature of the covariates used for modelling, makes the findings applicable to most settings where ART is administered among the HIV+ terminal patients. All the independent factors identified as significant factors affecting the survival time of HIV+ terminal patients in this study should be accurately recorded, tested, analysed and should be monitored closely in order to lengthen the survival time and to improve the wellness of the HIV+ terminal patients.

Some noticeable improvements in terms of immunologic, virologic and weight recoveries among the HIV/AIDS patients as reflected in quantile plots, suggest concerted effort given by the health care staff to ensure a long and healthy life for all South Africans as enshrined in the mission statement of the Department of Health. However, most Antiretroviral drugs (ARVs) have neuropsychiatric side-effects, for example efavirenz (EFV) is commonly associated with insomnia and headache (Reid et al., 2012). Awoke et al. (2016) determined and compared the long-term response of patients on nevirapine (NVP) and efavirenz (EFV) based first line antiretroviral therapy regimen in Ethiopia, and they found out that the hazard of composite outcome on nevirapine relative to efavirenz was more. In addition, ART may lead to the manifestation of TDF nephrotoxicity as has been mentioned above.

Historically, health information systems in South Africa have been characterised by fragmentation and lack of coordination, prevalence of manual systems and lack of automation (Department of Health , 2012). Some of these historical problems which tend to affect HIV/AIDS patients need to be fully addressed through the eHealth as outlined in the National eHealth Strategy of the period 2012/13-2016/2017. World Health Organisation as in Department of Health (2012), defines eHealth as the use of information and communication technologies (ICTs) for health to treat patients, pursue research, educate students, track diseases and monitor public health. If eHealth interventions are fully applied to HIV/AIDS programmes in South Africa, some further improvements in terms of ART adherence, lost to follow-up, transfers, mortality, data analysis and records keeping and retrieval would be enhanced through:

- Electronic Health Records (enabling sharing of patient data between points of care),
- Vital Registration (the use of computerised systems for registration of death or transfers or lost to follow-up),
- mHealth (for example use of mobile devices such as cell-phones to share information, to trace transferred patients or patients lost to follow-up, or to collect patient data),
- Telemedicine (for example, using of ICTs to provide care to patients at a distance), and
- Health Research (for example, using high-performance computing to handle large volumes of data).

Finally, Cox regression, logistic regression and quantile regression are compared in answering the research question of this study. Logistic regression analysis was performed not to directly

answer the question but as an exploratory analysis to determine whether Patient status (censored, uncensored) was associated with the covariates. The hypothesis was that if so, those covariates found to be associated with Patient status would also have an effect of the survival times of patients. This expectation was to some extent met as some but not all the significant covariates from the logistic modelling were significant in the Cox PH model. However, the Cox PH modelling should be preferred (versus the logistic modelling) as it directly answered the research question, and in terms of the number of significant factors, obtained one more than the logistic modelling. In addition, Cox model should be preferred over the logistic model when survival time information is available and there is censoring. As pointed out by Kleinbaum and Klein (2012), the Cox regression uses more information in form of 'survival times', whereas the logistic regression considers a (0, 1) outcome and ignores survival times and censoring.

The Cox PH modelling and the quantile regression analysis complemented each other in answering the research question. However, although the Cox PH modelling was the main approach in this study, the quantile regression analysis results are more informative than the Cox PH modelling results. With quantile regression, factors affecting each of the three different survival time quantiles (0.1 = short, 0.5 = median, 0.9 = long) where identified whereas with Cox regression only the factors affecting the "mean survival times" were identified. Hence, quantile regression modelling maybe more preferred for obtaining more informative results than Cox PH modelling. Comparisons of the two approaches in terms of both efficiency and robustness of identifying factors affecting the survival times can only be made for central survival times, and this is difficult. However, if Cox regression is regarded as "mean regression" theoretical mean regression is more efficient but less robust than median regression. Hence, the quantile regression approach should be recommended for modelling data suspected to have outliers.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

This chapter summarizes the study findings based on the objectives, the problem statement and the research question of the study and presents recommendations and suggestions for further research.

5.1 Conclusion

The study identified and modelled the factors affecting the survival of HIV+ terminal patients in Albert Luthuli Municipality by using Logistic regression, Cox PH regression, and Quantile regression modelling. For each statistical tool, survival time was regressed on independent factors which affect survival. Kaplan-Meier functions and Log-Rank test were used in statistical analysis of the stratified covariates. Hazard ratios and odds ratios were used as epidemiological measures of effect in Cox regression and Logistic regression respectively. Cox regression modelled the factors affecting the survival lifetime of HIV+ terminal patients as: ART adherence, Age, Follow-up mass, Baseline sodium, Baseline viral load and interactions of Follow-up lymphocyte by TB history and Follow-up CD4 by Treatment (Regimen 1). Logistic regression modelled the factors significantly associated with mortality among the HIV+ terminal patients as: Follow-up CD4; Marital status; Follow-up lymphocyte; Baseline viral load; Follow-up Alanine Transaminase and the interaction of Follow-up mass by Treatment (Regimen 1).

Quantile regression determined significant factors on the 0.1^{st} quantile of the log (survival time) as: poor and fair ART adherence relative to good ART adherence, *ln* (Follow-up viral load), Follow-up white blood cell count, Follow-up CD4 and the interaction effect of Follow-up white blood cell count by Treatment (Regimen 1). The significant factors with significant effects on the 0.5^{th} quantile of the log (survival time) at the 0.1 level of significance were determined as: WHO stage 3 relative to WHO stage 4, poor and fair ART adherence relative to good ART adherence and Baseline haemoglobin, Gender, Hospital, WHO stage 1 relative to WHO stage 4 and Follow-up CD4. Lastly, the significant factors with significant effects on the 0.9^{th} quantile of the log (survival time) at the 0.1 level of significance were determined as: poor and fair ART adherence relative to good ART adherence, ln (Baseline viral load), Carolina hospital

relative to Embhuleni hospital, Follow-up CD4, Baseline white blood cell count and Follow-up sodium.

5.2 Recommendations

Considering the findings in this study, recommendations are hereby made:

i) High patient transfer rates in Albert Luthuli Municipality warrants detailed documentation for transfer-outs and transfer-ins in order to keep track of national outcomes without double counting of transfer-outs and without some transfer-outs immediately turning into lost to follow-up. High patient lost to follow-up rate could be handled as in similar studies done by Sieleunou et al. (2008) in Far-North Province of Cameroon, whereby patients lost to follow-up were traced using either a cell-phone or messengers. In addition, Electronic Health Records as outlined in eHealth must be kept update to enable sharing of patient data between points of patient care.

ii) Monitoring of the patients under treatment could be improved for patients coming from remote areas (for example farms) to be adherent to ART. In addition, ART adherence and patient retention may be strengthened by having patients accompanied by their relatives for counselling sessions as in similar studies by Sieleunou et al. (2008). Telemedicine intervention strategies, as outlined under eHealth (Department of Health , 2012), may also be used to provide care to patients at a distance.

iii) High patient transfer and lost to follow-up rates in Albert Luthuli stand as indicators that retention mechanisms must be strengthened as was raised in similar studies by Alemu et al. (2010). The monitoring frequency of these outcomes may have to be raised from say quarterly to monthly or even higher depending on the seriousness of the outcomes.

iv) Data management in terms of recording system, records keeping, and retrieval may need improvement. The missing component in most South African hospitals is data analysis. If data capturers in hospitals could be supplemented with data analysts; that will empower Health workers and patients by availing most needed information at the right time.

v) Adherence to ART as a significant factor in all models in this study and as one of the key factors underlying treatment compliance may be strengthened through widespread use treatment literacy training of the HIV/AIDS patients. As pointed out by Karim et al (2010), treatment literacy training may be delivered at HIV/AIDS clinics or through community

outreach programme. The delivery could be done by lay or peer counsellors, lay HIV/AIDS educators or by other patients with good ART adherence.

vi) The factors which were found to be affecting the survival lifetime of HIV+ terminal patients which include ART adherence, Baseline viral load, Follow-up CD4 and Treatment (Regimen 1) need meticulous testing, recording and a close follow-up.

vii) The study established diverse baseline statistics against which future research could be done in other districts and provinces of South Africa for cross-comparison of the findings. If this study is collated with similar studies to be done here in South Africa, then the prognostic models produced could be used by health practitioners, researchers or by policy makers as predictive tools for predicting HIV/AIDS trends country wide. Once this is done, appropriate policies can be developed and be recommended for adoption at national level. According to Teka et al. (2015) who did similar studies; understanding the survival time of patients and its patterns is the key for policy formulation in bringing improvements.

Lastly, further research could be conducted in the following areas:

i) By considering all the previous studies on survival analysis cited in this research; it appears the length of follow-up time is a haphazard arrangement which is not informed by any findings or by any literature. That is, there is little theoretical or empirical evidence to use as a guide for the choice of the length of the follow-up time. This is reinforced by Perrigot et al. (2004) who suggested that the length of the observation period is a personal and arbitrary judgement by the researcher. This is even though the interpretation of study results can vary with the length of the observation period. Further research could be undertaken to bring some conformity in the length of follow-up time. Such a study would also establish the significant determinants of the length of follow-up time.

ii) While ART is associated with some improvement in virology and immunology in HIV-infected patients, research is still needed for the assessment of the impact of ART on renal and liver functions in marginalized communities in Africa for example in Albert Luthuli Municipality of South Africa.

References

Abebe, N., Alemu, K., Asfaw, T., & Abajobir, A.A. (2014). Survival status of hiv positive adults on antiretroviral treatment in Debre Markos. *Pan African Medical Journal 19(1)* Retrieved May 13, 2018, from http://www.panafrican-med-journal.com/content/article/17/88/full/

Alemu, A.W., & Sebastian, M.S. (2010). Determinants of survival in adult HIV patients on antiretroviral therapy in Oromiyaa, Ethiopia. *Glob Health Action, 3: 5398*. Retrieved July 19, 2018 from DOI: 10.3402/gha.v3i0.5398

Ansin, E. (2015). An evaluation of the Cox-Snell residuals. *Master's thesis.* Retrieved March 21, 2018, from https://www.diva-portal.org/smash/get/diva2:826234/FULLTEXT01.pdf

Avert. (2017). *Global information and education on HIV and AIDS: HIV and AIDS in Eswatini.* Retrieved May 17, 2020, from https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/eswatini

Awoke,T., Worku, A., Kebede, Y., Kasim, A., Birlie, B., Braekers, R., Zuma, K., & Shkedy, Z. (2016). Modeling Outcomes of First-Line Antiretroviral Therapy and Rate of CD4 Counts Change among a Cohort of HIV/AIDS Patients in Ethiopia: A Retrospective Cohort Study. Gondar, Ethiopia. *PLOS ONE*. Retrieved May 12, 2019, from doi:10.1371/journal.pone.0168323

Azia, I.N., Mukumbang, F.C., & van Wyk, B. (2016). Barriers to adherence to antiretroviral treatment in a regional hospital in Vredenburg, Western Cape, South Africa. Western Cape, South Africa: *AOSIS.* Retrieved February 17, 2019, from https://sajhivmed.org.za/index.php/hivmed/article/view/476/875#CIT0004_476

Bakanda, C., Birungi, J., Mwesigwa, R., Nachega, J.B., Chan, K., Alexis Palmer, A., Ford, N & Mills, E.J. (2011). Survival of HIV infected adolescents on antiretroviral therapy in Uganda: Findings from a nationally representative cohort in Uganda. Uganda. *PLOS ONE*. Retrieved June 21, 2018, from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019261

Bellavia, A. (2015). Quantile Regression in Survival Analysis. *Karolinska Institutet, Stockholm: Unit of Biostatistics, Institute of Environmental Medicine.* Retrieved February 19, 2018, from https://andreabellavia.github.io/bicocca.pdf

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 11: Assessing risk. *BioMed Central Ltd, 8, 287-291*. Retrieved April 11, 2019, from doi:DOI 10.1186/cc2908

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 13: Receiver operating. Crit Care. *BioMed Central Ltd, 9*, 508-512. Retrieved May 10, 2019, from doi:DOI 10.1186/cc3045

Bezuidenhout, S., Ogunsanwo, D.A., & Helberg, E.A. (2014). Patient satisfaction at accredited antiretroviral treatment sites in the Gert Sibande District. *Afr J Prm Health Care Fam Med, 6(1)*. Retrieved January 22, 2018, from doi:http://dx.doi.org/10.4102/phcfm.v6i1.627

Blundell, R., & Powell, J.L. (2007). Censored Regression Quantiles with Endogenous Regressors. *Journal of Econometrics*. Retrieved March 29, 2019, doi:10.1016/j.jeconom.2007.01.016

Boswell, M.T., & Rossouw, T.M. (2017). Approach to acute kidney injury in HIV-infected patients in South Africa. South Africa: *Southern African Journal of HIV Medicine*. Retrieved October 28, 2018, from doi:https://doi.org/10.4102/

Box-Steffensmeier, J.M., & Jones, B.S. (2004). *Event history modelling: a guide for social scientists.* Cambridge: University Press

Braconnier, P., Delforge, M., Maria Garjau, M., Wissing, K.M., & Stéphane De Wit, S. (2017). Hyponatremia is a marker of disease severity in HIV-infected patients: a retrospective cohort study. *BMC Infectious Diseases*. Retrieved May 23, 2019, from doi:10.1186/s12879-017-2191-5

Budziera, A., & Flyvbjerga, B. (2013). *Making-Sense of the Impact and Importance of Outliers in Project.* University of Oxford. Retrieved March 22, 2019, from eureka.sbs.ox.ac.uk/4745/1/Budzier_and_Flyvbjerg.pdf

Canette, I. (2014). Using resampling methods to detect influential points. Not elsewhere classified. *The STATA Blog.* Retrieved March 30, 2019, from https://blog.stata.com/2014/05/08/using-resampling-methods-to-detect-influential-points/

Chan, P. (2017). "*Measures of influence".* (2017, January 22). *Influential points-Cook's D, DEFFITS, DFBETAS.* [Video file]. Retrieved February 15, 2019, from https://youtu.be/31xA3hsxW6k

Chatterjee, D., & Chatterjee A. (2010). Binary Logistic Regression Using Survival Analysis. *SSRN Electronic Journal*. Retrieved March 12, 2019, from doi:DOI: 10.2139/ssrn.1672759

Chatterjee, D., & Hadi, S. (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science, Vol. 1, No. 3, 379-416.* Retrieved April 19, 2019, from https://www.jstor.org/stable/2245477

Chaudhuri, P. K. (1997). On average derivative quantile regression. *The Annals of Statistics, 25(2): 715-744.* Retrieved March 12, 2019, from https://www.jstor.org/stable/2242564

Chief Albert Luthuli Municipality. (2018). *Integrated Development Plan (IDP) 2018/2019.* Retrieved January 12, 2018, from https://cogta.mpg.gov.za/IDP/2018-19IDPs/Gert%20Sibande/ChiefAlbertLuthuli2018-19.pdf

Chow, S., Shao, J., & Wang, H. (2008). *Sample Size Calculations in Clinical Research*. 2nd Ed. Chapman & Hall/CRC Biostatistics Series.

Chu, K., Moyo, S., Ogunmefun, C., Mbatha, T., Bock, P., & English, R. (2011). *District Hospital Performance Assessment Mpumalanga Province 2008-2010.* Retrieved September 29, 2018, from http://www.hst.org.za

Clark, T.G., Bradburn, M.J., Love, S.B., & Altman, D.G. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer.* Retrieved May 23, 2017, from www.slaop.org/pdf/814Journ7.pdf

Cleves, M., Gould, W.W., & Marchenko, Y.V., (2010). *An introduction to Survival Analysis Using Stata.* University of Arkansas: Stata press

Cochran, W.G. (1977). *Sampling techniques (3rd edition).* New York. John Wiley & Sons.

Corkery, S. (2016). Diagnosed with HIV at a low CD4 count. *Nam Aidsmap HIV&AIDS sharing knowledge, changing lives.* Retrived July 11, 2018, from https://www.aidsmap.com/about-hiv/diagnosed-hiv-low-cd4-count

Cox, D. (1972). Regression Models and life-tables (with Discussion). *Joumal of the Royal Statistical Society Series, (Vols. B 34, 187-220.).* Retrieved May 16, 2018, from www.stat.cmu.edu/~ryantibs/journalclub/cox_1972.pdf

Croxford, S., Kitching, A., Desai, S., Kall, M., Edelstein, M., Skingsley, A. … Delpech, V. (2016). Mortality and causes of death in people diagnosed with HIV in the era of highly active antiretroviral therapy compared with the general population: an analysis of a national observational cohort. *Lancet Public Health 2017*. London, UK: Elsevier Ltd. Retrieved March 16, 2019, from http://dx.doi.org/10.1016/S2468-2667(16)30020-2

Crum-Cianflone, N., Ganesan, A., Teneza-Mora, N., Riddle, M., Medina, S., Barahona, I., & Brodine, S. (2010). Prevalence and factors associated with renal dysfunction among HIV-infected patients. *AIDS Patient Care STDS, 24(6):353-60.* Retrieved October 30, 2018, from doi:10.1 089/apc.2009.0326

Damtew, B., Mengistie, B., & Alemayehu, T. (2015). Survival and determinants of mortality in adult HIV/Aids patients initiating antiretroviral therapy in Somali Region, Eastern Ethiopia. *Pan African Medical Journal*, 22: 138. Retrieved March 12, 2018, from doi: 10.11604/pamj.2015.22.138.4352

DAura, T. *"How to Random Sample in Excel"*. (2013, January 19). How to Create a Random Sample in Excel (in 3 minutes!). [Video file]. Retrieved March 17, 2017, from https://youtu.be/q8fU001P2lI

Dayton, C. M. (1992). Logistic regression analysis. *(D. o. University of Maryland, Ed.) ResearchGate*. Retrieved April 21, 2019, from https://www.researchgate.net/publication/268416984

Department of Health. (2012). *eHealth Strategy South Africa 2012/13-2016/17*. Retrieved May 16, 2019, from https://ehna.acfee.org/e04574669772103243c5d4570a847fe716655b92.pdf

Desai, K.T., Joshi, N., Verma, A., Patel, P.B., & Bansal, R. (2015). Survival Analysis of HIV Positive Patients taking Anti-Retroviral Therapy under National AIDS Control Program in Gujarat. *International Journal of Epidemiology, 2015, 44, Supplement 1*. Retrieved May 15, 2018, from DOI: 10.1093/ije/dyv096.564

Eckel, S. (2008). Lecture 14: Interpreting logistic regression models. Retrieved May 18, 2019, from www-hsc.usc.edu/~eckel/biostat2/slides/lecture14.pdf

Electoral Commission of South Africa (IEC). (2016). Chief Albert Luthuli. *Municipal election results*. Retrieved May 17, 2020, from https://wazimap.co.za/profiles/municipality-MP301-chief-albert-luthuli/

El-Habil, A. (2012). An Application on Multinomial Logistic Regression Model. *Pakistan Journal of Statistics and Operation Research*. Retrieved April 12, 2019, from DOI: 10.18187/pjsor.v8i2.234

Eneyew, K. , Seifu, D. , Amogne, W., & Menon, M. (2016). Assessment of Renal Function among HIV-Infected Patients on Combination Antiretroviral Therapy at Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia. *Technology and Investment*, 7, 107-122. Retrieved November 5, 2018, from doi: 10.4236/ti.2016.73013

Estrella, M.M., & Fine, D.M. (2010). Screening for Chronic Kidney Disease in HIV-Infected Patients. *Adv Chronic Kidney Dis, 17(1):26-35.* Retrieved October 20, 2018, from doi:10.1053/j.ackd.2009.07.014

Etikan, I., & Babatope, G. (2018). Survival Analysis: A Major Decision Technique in Healthcare Practices. *International Journal of Science and Research Methodology, Vol. 8 (4): 121-135. Retrieved July 25, 2018, from* http://ijsrm.humanjournals.com/wp-content/uploads/2018/03/12.%C4%B0lker-Etikan-Ogunjesa-Babatope.pdf

Etikan, I., Abubakar, S & Alkassim, R. (2017). The Kaplan Meier Estimate in Survival Analysis. *Biom Biostat Int J 5(2): 00128.* Retrieved January 21, 2018, from http://medcraveonline.com/BBIJ/BBIJ-05-00128.php

Fay, M.P., Brittain, E.H., & Proschan, M.A. (2013). Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics (2013),* ***14*** *(4)723–736*. Retrieved June 19, 2020, from doi:10.1093/biostatistics/kxt016

Fletcher, J. (2018). What is an HIV viral load? Medical news today. Retrieved Retrieved May 8, 2019, from https://www.medicalnewstoday.com/articles/323851.php

Flom, P.L., & Peter Flom Consulting. (2011). Quantile regression with PROC QUANTREG. New York, USA: *NESUG 2011 Statistics & Analysis.* Retrieved March 15, 2019, from https://www.lexjansen.com/nesug/nesug11/sa/sa04.pdf

Giancristofaro, R.A., & Salmaso, L. (2003). Model performance analysis and model validation in logistic regression. *STATISTICA, anno LXIII, n. 2*. Retrieved March 17, 2019, from

https://rivista-statistica.unibo.it/article/download/358/351

Goel, M.K., Khanna, P., & Kishore , J. (2010). Understanding survival analysis. *International Journal of Ayurveda Research, Vol 1*(4). Retrieved November 30, 2017, from doi:10.4103/0974-7788.76794

Gorfine, M., Goldberg, Y., & Ritov, Y. (2014). A quantile regression model for failure-time data with time-dependent covariates. *Biostatistics*. July 15, 2018, from doi:10.1093/biostatistics/kxw036

Gorfine, M., Goldberg, Y., & Ritov, Y. (2017). A quantile regression model for failure-time data with time-dependent covariates. Retrieved April 19, 2019, from https://www.ncbi.nlm.nih.gov/pubmed/27485534

Goulder, P.J ., Lewin, S.R ., & Leitman, E.M. (2016). Paediatric HIV infection: the potential for cure. *Nat Rev Immunol. 2016 April, 16(4): 259–271*. Retrieved June 12, 2020, from doi:10.1038/nri.2016.19

Guffey, D. (2012). Hosmer-Lemeshow goodness-of-fit test: Translations to the Cox Proportional Hazards Model. *Master's thesis.* University of Washington. Retrieved March 21, 2019, from https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/22648/Guffey_washington_0250O_11143.pdf;sequence=1.

Hamidi, O., Poorolajal, J., & Tapak, L. (2017). Identifying predictors of progression to AIDS and mortality post-HIV infection using parametric multistate model. *Epidemiology Biostatistics and Public Health, 14(2).* Retrieved May 20, 2020, from *DOI: 10.2427/12438*

Harrell Jr, F. (2018). Regression Modeling Strategies. *Biostatistics for Biomedical Research*. Nashville, USA: Vanderbilt University School of Medicine. Retrieved November 11, 2018, from biostat.mc.vanderbilt.edu/rms

Hosmer Jr, D.W., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data: (Second Edition ed.)*. New York, NY. John Wiley and Sons Inc

Huang, B., Yu, C.-R., & Chuang-Stein, C. (2013). Kaplan-Meier Estimator, Statistics2013. [Poster]. Retrieved April 12, 2018, from https://www.slideserve.com/morse/kaplan-meier-estimator

Human body weight. (2018, June 17). In *Wikipedia.* Retrieved June 17, 2018, from https://en.m.wikipedia.org/wiki/Human_body_weight

Hurlin, C. (2015). Advanced Econometrics-Master ESA. Chapter 4: Statistical Hypothesis Testing. *University of Orleans.* Retrieved April 30, 2018, from https://www.univ-orleans.fr/deg/masters/ESA/CH/Chapter4_Inference.pdf

HyLown Consulting LLC. (2020). *Calculate Sample Size Needed to Test Time-To-Event Data: Cox PH, Equivalence.* Retrieved June 12, 2020, from http://powerandsamplesize.com/Calculators/Test-Time-To-Event-Data/Cox-PH-Equivalence

Ijomah, A., & Nwali, O. (2018). A comparative study of some variable selection techniques in Logistic regression. *European Journal of Mathematics and Computer Science, 5(1)*. Retrieved May 13, 2019, from https://www.idpublications.org/.../Full-Paper-A-comparative-study-of-some-...

Johnson, L.L., & Shih J.H. (2012). An Introduction to Survival Analysis in Principles and Practice of Clinical Research. Retrieved May 27, 2018, from https://www.sciencedirect.com/topics/neuroscience/survival-analysis

Kalbfleisch, J.D., & Prentice, R.L. (2002). The Statistical Analysis of Failure Time Data. *John Wiley & Sons, Inc.* Retrieved April 20, 2018, from https://pdfs.semanticscholar.org/7624/103ae633517b78510ee6619168ccdab881cd.pdf

Karadeniz, P.G., & Ercan, I. (2017). Examining tests for comparing survival curves with right censored data. *STATISTICS IN TRANSITION new series, 18(2),311–328*. Retrieved March 26, 2018, from https://stat.gov.pl/download/gfx/portalinformacyjny/en/defaultlistaplikow/3503/15/1/examining_tests_for_comparing_survival_curves_with_right_censored_data.pdf

Karim S.S.A. (2014). State of the Art: Epidemiology and Access. *UNAIDS, CAPRISA.* Retrieved June 23, 2018, from http://strive.lshtm.ac.uk/sites/strive.lshtm.ac.uk/files/State%20of%20the%20Art-Epidemiology%20and%20Access.pdf

Karim, S.S.A. (2014). The HIV Epidemic: Progress & Challenges Southern African HIV Clinicians Society Conference. *UNAIDS Global Report*. Retrieved April 17, 2018, from https://sahivsoc.org/Files/Fri_Salim_Karim-the-HIV-Epidemic-Progress-and-challenges.pdf

Karim, S.S.A., Churchyard, G.J., Karim, Q.A., & Lawn, S.D. (2010). HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. *NIH Public Access 374(9693): 921–933.* Retrieved May 12, 2018, from doi:10.1016/S0140-6736(09)60916-8

Kim, S & Kim H. (2014). Relative lymphocyte count as a marker of progression of chronic kidney disease. *PubMed, International Urology and Nephrology 46(7)*. doi:10.1007/s11255-014-0687-0

Klein, J.P & Moeschberger, M.L. (2003). Survival analysis Techniques for Censored and Truncated Data. Ohio. USA. *Springer.*

Kleinbaum, D.G., & Klein, M. (2012). Survival Analysis. A Self-Learning Text. *Springer*. Retrieved May 16, 2020, from DOI: 10.1007/978-1-4419-6646-9, pdf

Koenker, R., & Machado, J.A. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association, 1296-1310.* Retrieved May 7, 2019, from doi:http://dx.doi.org/10.1080/01621459.1999.10473882

Kwong, G.P.S., & Hutton, J.L. (2003). Choice of parametric models in survival analysis: Applications to monotherapy for epilepsy and cerebral palsy. *Applied Statistics 52(2):153-168.* Retrieved June 12, 2020, from DOI: 10.1111/1467-9876.00395

Leng, C., & Tong, X. (2013). A quantile regression estimator for censored data. *Bernoulli 19(1), 344–361*. Retrieved January 14, 2019, from https://arxiv.org/pdf/1302.0181

Li, J., & Ma, S. (2013). *Survival analysis in medicine and genetics.* Broken Sound Parkway NW, US. CRC Press, Taylor and Francis Group, LLC.

Liddle, A. (2008). Information criteria for astrophysical model selection. *(U. o. Astronomy Centre, Ed.) UK.* Retrieved April 27, 2019, from http:/ / arxiv. org/ PS_cache/ astro-ph/ pdf/ 0701/ 0701113v2. pdf

Lin, D.Y., Wei, L.J., & Ying, Z. (2002). Model-Checking Techniques Based on Cumulative Residuals. *Biometrika, 58, 1-12.* Retrieved April 17, 2019, from https://www.jstor.org/stable/3068284

Lin, G., Rodriguez, R.N. (2013). Using the QUANTLIFE Procedure for Survival Analysis. *SAS Global Forum, Statistics and Data Analysis, (Paper 421-2013). SAS Institute Inc.* Retrieved July 13, 2018, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.306.7123&rep=rep1&type=pdf

Luz, P. M., Girouard, B., Grinsztejn, K. A., Freedberg, V. G., Veloso, E., Losina, C. J. … Walensky, R.P. (2016). Survival benefits of antiretroviral therapy in Brazil: a model-based. *Journal of the International AIDS Society*. Retrieved April 19, 2018, from http://www.jiasociety.org/index.php/jias/article/view/20623

Mack, C.A. "*Lecture 22 Influence in Regression".* (2016, September 16). *Lecture22(Data2Decision) Influence in Regression*. [Video file]. Retrieved February 15, 2019, from https://youtu.be/DUd_soPQf1c

Maposa, I. (2016). Survival modelling and analysis of HIV/AIDS patients on HIV care and antiretroviral treatment to determine longevity prognostic factors. *Master's thesis*. Retrieved September 13, 2017, from https://pdfs.semanticscholar.org/826a/0be64365d309d5d6ebe1c5bc72699978cfa3.pdf

Mashinya, F., Alberts, M., Colebunders, R., Van Geertruyden, J-P. (2016). Weight status and associated factors among HIV-infected people on antiretroviral therapy in rural Dikgale, Limpopo, South Africa. *African Journal of Primary Health Care & Family Medicine, 8(1).* Retrieved November 5, 2018, from DOI: 10.4102/phcfm.v8i1.1230

Masinga, S. (2014, April 4), Inside South Africa's HIV and AIDS capital Gert Sibande District. *Ziwaphi.* Retrieved from https://www.ziwaphi.com/2014/04/04/inside-south-africas-hiv-and-aids-capital-gert-sibande-district/ on 14 May 2020

May, M., Sterne, J., & Egger, M. (2003). Parametric survival models may be more accurate than Kaplan-Meier estimates. *British Medical Journal*, 326(7393): 822. Retrieved May 18, 2020, from **doi:** 10.1136/bmj.326.7393.822

McGready, J. (2009). Regression for Survival Analysis. *Baltimore, USA: Johns Hopkins University.* Retrieved January 08, 2019, from ocw.jhsph.edu/courses/StatisticalReasoning2/PDFs/2009/StatR2_lec10a_mcgready.pdf

McManus, H., O'Connor, C.C., Boyd, M., Broom, J., Russell, D.,Watson, K., Roth, N., Read, P.J., Petoumenos, K & Lawet, M.G. (2012). Long-Term Survival in HIV Positive Patients with up to 15 Years of Antiretroviral Therapy. (N. I. Michael Alan Polis, Ed.) *PLOS ONE*. Retrieved June 26, 2018, from doi:10.1371/journal.pone.0048839

Melnyk, A., Pagell, M and Jorae, G. (1995). Applying survival analysis to operations management: Analyzing the differences in donor classes in the blood donation process. *Journal of Operations Management 13 (1995) 339-356.* Retrieved May 15, 2020, from https://www.sciencedirect.com/science/article/pii/0272696395000313

Merab, E. (2018). Why men are more likely to die of HIV. *Nation Media Group. Nairobi: Nairobi News.* Retrieved April 30, 2019 from https://nairobinews.nation.co.ke/news/why-men-are-more-likely-to-die-of-hiv

Ministry of Health and Social Services, Directorate of Special Programmes. (2010). *National Guidelines for Antiretroviral Therapy (Fourth Edition).* Retrieved June 17, 2018, https://aidsfree.usaid.gov/sites/default/files/tx_namibia_2013.pdf

Mlangeni, N & Senkubuge, F. (2016). Antiretroviral therapy programme outcomes in Tshwane district, South Africa: A 5-year retrospective study. *South African Medical Journal, 106(4):365-368.* Retrieved June 24, 2018, from DOI:10.7196/SAMJ.2016.v106i4.9375

Moshago, T., Haile, D & Enkusilasie, F. (2014). Survival Analysis of HIV Infected People on Antiretroviral Therapy at Mizan-Aman General Hospital, Southwest Ethiopia. *International Journal of Science and Research (IJSR), 3(5).* Retrieved March 25, 2018, from http://www.ijsr.net/archive/v3i5/MDIwMTMyMTAz.pdf

Motsoaledi, A. (2013, November 22). Mpumalanga's Gert Sibande district has highest HIV rate. *City Press.* Retrieved April 12, 2017, from http://www.news24.com/Archives/City-Press/Mpumalangas-Gert-Sibande

Mpande, B. (2018, December 8). Teenage pregnancy on the rise. *Lowvelder.* Retrieved May 16, 2020, from https://lowvelder.co.za/462611/teenage-pregnancy-rise/

Mpumalanga Provincial AIDS Council. (2016). *Annual progress report 2014/15.* Retrieved

December 10, 2017, from http://sanac.org.za/download/563/resources/3385/mp_psp-annual-progress-report_final-report.pdf

Mpumalanga Department of Health. (2016). *District Health Information System (DHIS) Data; 2015/16.* . Retrieved March 13, 2018, from sanac.org.za/wp.../2016/.../MP_PSP-ANNUAL-PROGRESS-REPORT_Final-Report.p on

Mpumalanga Municipalities. (2017). *Chief Albert Luthuli Local Municipality (MP301)*. Retrieved March 19, 2017 from https://municipalities.co.za/locals/view/147/Chief-Albert-Luthuli-Local-Municipality#demographic

Muenchhoff, M., Adland, E., Karimanzira, O., Crowther, C., Pace, M., Anna Csala, A., …, Goulder, P. (2016). Non-progressing HIV-infected children share fundamental immunological features of non-pathogenic SIV infection. *Sci Transl Med. 8(358): 358ra125.* Retrieved June 12, 2020 from doi:10.1126/scitranslmed.aag1048

Nagu , T.J., Kanyangarara, M., Hawkins, C., Hertmark, E., Chalamila, G., Spiegelman, D., Mugusi, F., & Fawzi, W. (2012). Elevated alanine aminotransferase in antiretroviral-naïve HIVinfected African infected patients: magnitude and risk factors. Dar es Salaam, Tanzania: *NIH Public Access*. doi:10.1111/j.1468-1293.2012.01006.x

Nakhaee, F & Law, M. (2011). Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy: data from Australia. *Eastern Mediterranean Health Journal, 17(3).* Retrieved April 17, 2018, from https://pdfs.semanticscholar.org/0046/d27a30da17665b5075263cd2b62b952206ae.pdf

National Department of Health. (2015). *The 2015 National Antenatal Sentinel HIV & Syphilis Survey, South Africa, National Department of Health.* Retrieved May 20, 2019, from http://www.health.gov.za/index.php/shortcodes/download=2584:2015-national-antenatal-hiv-prevalence-survey-final-23oct17

Nearly 6 000 babies born to Mpumalanga schoolgirls in 2017. (2018, August 16). *Independent*

Online. https://www.iol.co.za/the-star/nearly-6-000-babies-born-to-mpumalanga-schoolgirls-in-2017-16597952

Newson, R. (2010). Comparing the predictive powers of survival models using Harrell's C or Somers' D. *Stata Journal, StataCorp LP, 10(3)*:1-20. Retrieved April 29, 2019, from DOI: 10.22004/ag.econ.159022

Nkosi, J. (2017). *Chief Albert Luthuli local municipality Integrated Development Plan (IDP) 2017-2022.* Caroline. Mpumalanga. Chief Albert Luthuli Local Municipality

Ochieng-Ooko, V., Ochieng D., Sidle, J.E., Holdsworth, M., Wools-Kaloustian, K., Siika, A.M., Yiannoutsos, C.T., Owiti, M., Kimaiyo, S., & Braitstein P. (2010). Influence of gender on loss to follow-up in a large HIV treatment programme in western Kenya. *Bull World Health Organ, 88(9): 681–688.* Retrieved April 7, 2019, from doi: 10.2471/BLT.09.064329

Ogunsanwo, D.A. (2014). Determination of patient satisfaction at accredited antiretroviral treatment sites in the Gert Sibande District, Mpumalanga , South Africa. *Master's thesis.* University of Limpopo. Retrieved November 15, 2017, from http://ulspace.ul.ac.za/bitstream/handle/10386/778/Damilola%20dissertation%2019042012.pdf?sequence=1&isAllowed=y

Peng, L., & Huang, Y. (2008). Survival Analysis With Quantile Regression Models. *Journal of the American Statistical Association*. Retrieved February 28, 2019, from doi:10.1198/016214508000000355

Perrigot, R., Cliquet, G., & Mesbah, M. (2004). Possible applications of survival analysis in franchising research. *International Review of Retail, Distribution and Consumer Research, 14(1): 129-143.* Retrieved March 24, 2017, from http://www.tandfonline.com/doi/abs/10.108

Ranganathan, R., & Pramesh, C.S. (2012). Censoring in survival analysis: Potential for bias *PubMed, Perspectives in clinical research 3(1):40.* Accessed June 11, 2020, from

DOI: 10.4103/2229-3485.92307

Rao, C.M., & Rao, S.R. (2004). Sample Size Calculator. Retrieved January 19, 2017, from Raosoft, Inc.: www.raosoft.com/samplesize.html

Reid, E., Orrell, C., Stoloff, K., & Joska, J. (2012). Psychotropic prescribing in HIV. *Southern African Journal of HIV Medicine, 13(4).* Retrieved September 28, 2018, from September 28, DOI: 10.4102/sajhivmed.v13i4.115

Richardson, M. (2009). Investigation of Over-fitting and Optimism in Prognostic Models. *Thesis for the degree of Doctor of Philosophy*. School of Health and Population Sciences. The University of Birmingham. Retrieved March 18, 2019, from http://etheses.bham.ac.uk/754/1/Richardson10PhD.pdf

Riddlesworth, T. (2011). Estimation for the Cox model with various types of censored data. *Electronic Theses and Dissertations. 1705.* Retrieved from http://stars.library.ucf.edu/etd/1705

Rodriguez, R., & Yao, Y. (2017). Five Things you should know about quantile regression. *SAS Institute Inc, Paper SAS525-2017*. Retrieved January 15, 2018, from https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf

Ronan, F. "Sample size for survival analysis. A guide to planning successful clinical trials." (2018, Aug 29). Power and Sample Size Calculations for Survival Analysis - Webinar. Statsols, nQuery, Webnar on demand. [Webnar video file]. Retrieved June 12, 2020, from https://www.statsols.com/webinar/sample-size-for-survival-analysis-webinar-on-demand?submissionGuid=d3db594e-5cc2-4a69-b430-dcad60dd9f22

Saegusa, T., Di, C., & Chen, Y.Q. (2014). Hypothesis Testing for an Extended Cox Model with Time-Varying Coefficients. *Biometrics, 70(3)*: 619–628. doi:10.1111/biom.12185

Sanne, I.M., Westreich, D., Macphail, A.P., Rubel, D., Majuba, P & Van Rie A. (2009). Long

term outcomes of antiretroviral therapy in a large HIV/AIDS care clinic in urban South Africa: a prospective cohort study. *Journal of the International AIDS Society, 12:38.* Retrieved March 18, 2018, from doi:10.1186/1758-2652-12-38

Sarkar, S. K., Midi, H., & Rana, S. (2011). Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study. *Journal of Applied Sciences, 11*, 26-35. Retrieved February 19, 2019, from doi:10.3923/jas.2011.26.35

SAS Institute Inc. (2015). The QUANTREG Procedure. Retrieved February 20, 2019, from https://support.sas.com/documentation/onlinedoc/stat/141/qreg.pdf

Schmid, M., Wright, M.N., & Andreas Ziegler, A. (2016). On the use of Harrell's C for clinical risk prediction via random survival forests. *stat.ML*. Retrieved May 04, 2019, from https://arxiv.org/pdf/1507.03092

Schreiber-Gregory,D., & Jackson, H.M. Foundation. (2018). Logistic and Linear Regression Assumptions: Violation Recognition and Control. Retrieved May 02, 2019, from https://www.lexjansen.com › wuss › 2018 › 130_Final_Paper_pdf

Schulze, N. (2004). Applied Quantile Regression: Microeconometric, Financial, and Environmental Analyses. *Inaugural-Dissertation.* Universität Tübingen. Retrieved January 27, 2019, from https://publikationen.uni-tuebingen.de/xmlui/bitstream/handle/10900/47317/pdf/Schulze_-_Applied_Quantile_Regression.pdf?sequence=1

Shebeshi, D.S. (2011). Survival Analysis of adult HIV/AIDS patients and stochastic modelling of AIDS disease progression: A case study of Jimma University specialised hospital, Ethiopia. *(H. University, Ed.) Hawassa, Ethiopia*. Retrieved May 12, 2018, from DOI: 10.13140/2.1.3756.8323

Sieleunou, I., Souleymanou, M., Schönenberger, A.M., Menten, J., & Boelaert, M. (2008). Determinants of Survival in AIDS patients on antiretroviral therapy in a rural centre in the Far-North Province, Cameroon. *Tropical Medicine and International Health*,

14(1): 36–43. Retrieved April 17, 2018, from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3156.2008.02183.x/full

Slack, A., Yeoman, A., & Wendon, J. (2010). Renal dysfunction in chronic liver disease. London, UK: Institute of Liver Studies, King's College Hospital. *Crit Care, 14(2): 214.* Retrieved May 15, 2019, from doi: 10.1186/cc8855

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *Technical Report No. 4.* Retrieved May 23, 2018, from https://mafiadoc.com/dealing-with-missing-data-key-assumptions-and-methods-for-applied-_597b41be1723ddad8e079056.html

Statistics Solutions. (2019). Assumptions of Logistic Regression. *Advancement Through Clarity*. *Clearwater, USA.* Retrieved April 13, 2019, from https://www.statisticssolutions.com › uploads › wp-post-to-pdf-enhanced-cache

Statistics South Africa. (2011). *Census 2011 Municipal report.* Pretoria, South africa: Statistics South Africa. Retrieved June 13, 2017, from www.statssa.gov.za/census/census_2011/census_products/MP_Municipal_Report.pdf

Storvik, G. (2011). Numerical optimization of likelihoods:Additional literature for STK2120. Oslo: *University of Oslo*. Retrieved May 16, 2018, from https://www.mn.uio.no/math/tjenester/kunnskap/.../num_opti_likelihoods.pdf

Sullivan, L. (2016). Survival Analysis. *Boston University School of Public Health*. Retrieved March 15, 2018, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/...Survival/BS704_Survival_print.html

Tadege, M. (2018). Time to death predictors of HIV/AIDS infected patients on antiretroviral therapy in Ethiopia. Amhara, Ethiopia. *BMC Research Notes.* Retrieved April 12, 2019, from https://doi.org/10.1186/s13104-018-3863-y

Tadesse, K., Haile, F & Hiruy, N. (2014). Predictors of mortality among patients enrolled on antiretroviral therapy in Aksum Hospital, Northern Ethiopia: A Retrospective Cohort Study. *PLOS ONE, 9(1).* Retrieved May 6, 2018, from doi:10.1371/journal.pone.0087392

Tai, B.C., Chen, Z.J., & Machin, D. (2018). Estimating sample size in the presence of competing risks – Cause-specific hazard or cumulative incidence approach? *Statistical Methods in Medical Research, 2018, 27(1) 114–125.* Retrieved June 12, 2020, from DOI: 10.1177/0962280215623107

Teka, Z., Gizaw, Z., & Workneh, G. (2015). Investigating Correlates of the Survival of HIV/AIDS Patients Treated Under ART Follow-up: The Case of University of Gondar Hospital, Northwest Ethiopia. *Princetown University*. Retrieved May 28, 2019, from uaps2015.princeton.edu/abstracts/150168

The Pennsylvania State University. (2018). Identifying Influential Data Points. *Regression Methods*. *Eberly College of Science, Department of Statistics Online Programs*. Retrieved on March 28, 2019, from https://newonlinecourses.science.psu.edu/stat501/node/340/

Trickey, A., May, M.T., Vehreschild, J.-J., Obel, N., Gill, M.J., Crane, H.M., Boesecke, C. … Zangerle, R. (2017). Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. Retrieved May 24, 2018, from DOI:https://doi.org/10.1016/S2352-3018(17)30066-8

Tse, L. T. (1993). Application of Logistic regression and survival analysis to the study of CEP, manpower performance and attrition. *Master's thesis.* Singapore: BSinME.,National University of Singapore. Retrieved February 7, 2019, from https://apps.dtic.mil/dtic/tr/fulltext/u2/a273262.pdf

UNAIDS. (2016). HIV and AIDS in South Africa. *`Gap Report 2016'.* Geneva, Switzerland: UNAIDS. Retrieved May 23, 2017, from www.unaids.org/en/media/unaids/.../unaidspublication/.../UNAIDS_Gap_report_en.

pdf

University of California, Los Angeles, Statistical Consulting Group. (2018). *Introduction to SAS.* Retrieved March 4, 2018, from https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/

University of South Africa. (2017). University of South Africa Logo. *University of South Africa.* Retrieved from http://www.unisa.ac.za/sites/corporate/default/_

Veugelers, P.J., Cornelisse,P.G.A., Craib, K.J.P., Marion, S.A., Hogg, R.S., Strathdee, S.A., … Schechter, M.T. (1998). Models of survival in HIV infection and their use in the quantification of treatment benefits. *American Journal of Epidemiology*, 148: 487–496. Retrieved May 20, 2020, from https://pdfs.semanticscholar.org/8454/03026bfab710e34c91d6ecef0aa7d061b6d2.pdf

Viljamaa, S. (2017). Improved multivariate outlier removal in high volume IC production tests. *Master's Thesis.* University of Oulu Faculty of Technology. Retrieved February 11, 2019, from https://www.semanticscholar.org/paper/Improved-multivariate-outlier-removal-in-high-IC-Viljamaa/c89bd19a444ae765d8b1e80aab2983e6546d6267

Walpole, C., Prieto-Merino, D., Edwards, P., Cleland, J., Stevens, G., & Roberts, I. (2012). "The weight of nations: an estimation of adult human biomass". *BMC Public Health 2012, 12:439.* Retrieved August 19, 2017, from http://www.biomedcentral.com/1471-2458/12/439

Walters, S. (2012). Analyzing time to event outcomes with a Cox regression model. *Wiley Online Library, Wires Computational Statistics.* Retrieved May 21, 2018, from https://doi.org/10.1002/wics.1197

Wang, H.J., Zhou, J., & Li, Y. (2013). Variable selection for censored quantile regression. *Statistica Sinica, 23, 145-167.* Retrieved January, 2019, from doi:http://dx.doi.org/10.5705/ss.2011.100

Wilson, M.G. (2013). Assessing Model Adequacy in Proportional Hazards Regression. *SAS Global Forum 2013, Statistics and Data Analysis. . Indianapolis IN,, USA.* Retrieved July 13, 2018, from support.sas.com/resources/papers/proceedings13/431-2013.pdf

Windsperger, J., Cliquet, G., Hendrikse, G & Tuunanen, M. (2012). Economics and Management of Franchising Networks. *Springer Science & Business Media.* Retrieved July 23, 2018, from DOI 10.1007/978-3-7908-2662-3

Xu, R., Vaida, F., & Harrington, D.P. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica 19, 819-842.* Retrieved March 11, 2019, from http://www3.stat.sinica.edu.tw/statistica/oldpdf/A19n223.pdf

Yu, K., & Moyeed, R.A. (2001). Bayesian quantile regression. *Statistics & Probability Letters, 54, 437-447.* Retrieved January 17, 2019, from https://econpapers.repec.org/RePEc:eee:stapro:v:54:y:2001:i:4:p:437-447

Zhang, D. (2000). *4 Two (K) Sample Problems.* Retrieved February 27, 2018, from https://www4.stat.ncsu.edu/~dzhang2/st745/chap4.pdf

Zhang, H. (2015). Checking proportionality for Cox's regression model. *Master's thesis.* University of Oslo. Retrieved May 23, 2018 from https://www.duo.uio.no/bitstream/handle/10852/45324/HuiHongZhang_thesis.pdf?sequence=1&isAllowed=y

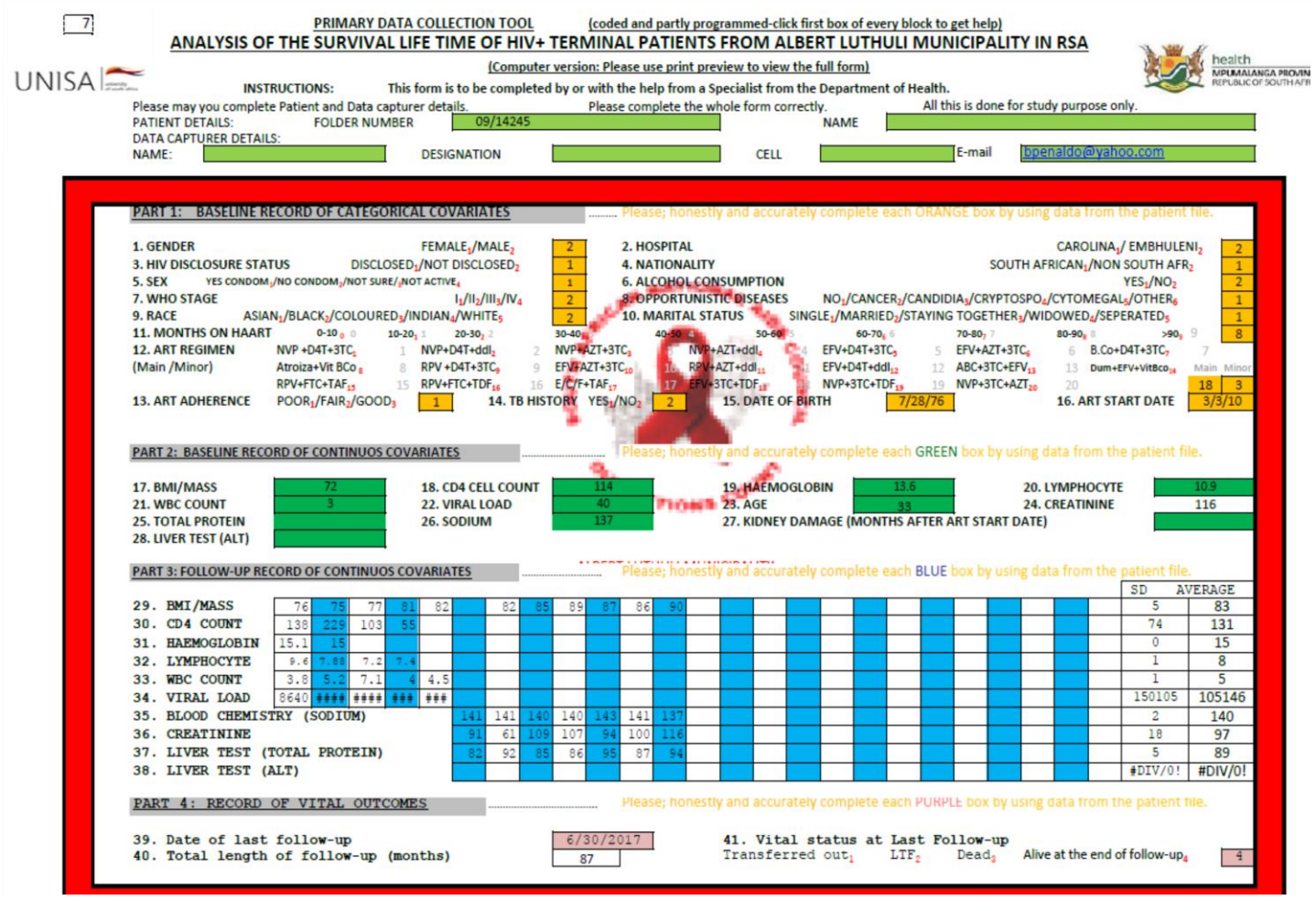
Zhao, F. (1998). Bootstrap variable selection and model validation for Cox's proportional hazards regression models - with applications to the identification of factors predictive of overall and post-relapse survival in advanced epithelial ovarian cancer. *Master's thesis.* Department of Community Health and Epidemiology. Queen's University Kingston. Retrieved March 02, 2019, from www.collectionscanada.gc.ca/obj/s4/f2/dsk2/tape17/PQDD_0026/MQ31275.pdf

Zhou, H. (2016). EM and MM Algorithms. *Los Angeles, USA: Department of Biostatistics University of California.* Retrieved June 10, 2018, from https://www.samsi.info/wp.../2016/08/Zhou_samsi-opt-summerschool-20160810-1.pd

Zhu et al. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statistica Sinica 22*, 1379-1401. Retrieved May 14, 2019, from doi:http://dx.doi.org/10.5705/ss.2010.199

Appendices

Appendix A: Research Data capturing tool (Excel)



7

PRIMARY DATA COLLECTION TOOL (coded and partly programmed-click first box of every block to get help)

ANALYSIS OF THE SURVIVAL LIFE TIME OF HIV+ TERMINAL PATIENTS FROM ALBERT LUTHULI MUNICIPALITY IN RSA

UNISA

health MPUMALANGA PROVIN REPUBLIC OF SOUTH AFR

(Computer version: Please use print preview to view the full form)

INSTRUCTIONS: This form is to be completed by or with the help from a Specialist from the Department of Health.

Please may you complete Patient and Data capturer details. Please complete the whole form correctly. All this is done for study purpose only.

PATIENT DETAILS: FOLDER NUMBER 09/14245 NAME

DATA CAPTURER DETAILS:

NAME: DESIGNATION CELL E-mail bpenaldo@yahoo.com

PART 1: BASELINE RECORD OF CATEGORICAL COVARIATES ..... Please; honestly and accurately complete each ORANGE box by using data from the patient file.

1. GENDER	FEMALE$_1$/MALE$_2$	2	2. HOSPITAL	CAROLINA$_1$/ EMBHULENI$_2$	2
3. HIV DISCLOSURE STATUS	DISCLOSED$_1$/NOT DISCLOSED$_2$	1	4. NATIONALITY	SOUTH AFRICAN$_1$/NON SOUTH AFR$_2$	1
5. SEX	YES CONDOM$_1$/NO CONDOM$_2$/NOT SURE/$_3$NOT ACTIVE$_4$	1	6. ALCOHOL CONSUMPTION	YES$_1$/NO$_2$	2
7. WHO STAGE	I$_1$/II$_2$/III$_3$/IV$_4$	2	8. OPPORTUNISTIC DISEASES	NO$_1$/CANCER$_2$/CANDIDIA$_3$/CRYPTOSPO$_4$/CYTOMEGAL$_5$/OTHER$_6$	1
9. RACE	ASIAN$_1$/BLACK$_2$/COLOURED$_3$/INDIAN$_4$/WHITE$_5$	2	10. MARITAL STATUS	SINGLE$_1$/MARRIED$_2$/STAYING TOGETHER$_3$/WIDOWED$_4$/SEPERATED$_5$	1
11. MONTHS ON HAART	0-10$_0$ 0 10-20$_1$ 1 20-30$_2$ 2 30-40$_3$ 3 40-50$_4$ 4 50-60$_5$ 5 60-70$_6$ 6 70-80$_7$ 7 80-90$_8$ 8 >90$_9$ 9				8
12. ART REGIMEN (Main /Minor)	NVP +D4T+3TC$_1$ 1; NVP+D4T+ddI$_2$ 2; NVP+AZT+3TC$_3$ 3; NVP+AZT+ddI$_4$ 4; EFV+D4T+3TC$_5$ 5; EFV+AZT+3TC$_6$ 6; B.Co+D4T+3TC$_7$ 7; Atroiza+Vit BCo$_8$ 8; RPV +D4T+3TC$_9$ 9; EFV+AZT+3TC$_{10}$ 10; RPV+AZT+ddI$_{11}$ 11; EFV+D4T+ddI$_{12}$ 12; ABC+3TC+EFV$_{13}$ 13; Dum+EFV+VitBco$_{14}$; RPV+FTC+TAF$_{15}$ 15; RPV+FTC+TDF$_{16}$ 16; E/C/F+TAF$_{17}$ 17; EFV+3TC+TDF$_{18}$ 18; NVP+3TC+TDF$_{19}$ 19; NVP+3TC+AZT$_{20}$ 20			Main 18 Minor 3	
13. ART ADHERENCE	POOR$_1$/FAIR$_2$/GOOD$_3$	1	14. TB HISTORY YES$_1$/NO$_2$	2	
15. DATE OF BIRTH	7/28/76		16. ART START DATE	3/3/10	

PART 2: BASELINE RECORD OF CONTINUOS COVARIATES ..... Please; honestly and accurately complete each GREEN box by using data from the patient file.

17. BMI/MASS	72	18. CD4 CELL COUNT	114	19. HAEMOGLOBIN	13.6	20. LYMPHOCYTE	10.9
21. WBC COUNT	3	22. VIRAL LOAD	40	23. AGE	33	24. CREATININE	116
25. TOTAL PROTEIN		26. SODIUM	137	27. KIDNEY DAMAGE (MONTHS AFTER ART START DATE)			
28. LIVER TEST (ALT)							

PART 3: FOLLOW-UP RECORD OF CONTINUOS COVARIATES ..... Please; honestly and accurately complete each BLUE box by using data from the patient file.

														SD	AVERAGE
29. BMI/MASS	76	75	77	81	82		82	85	89	87	86	90		5	83
30. CD4 COUNT	138	229	103	55										74	131
31. HAEMOGLOBIN	15.1	15												0	15
32. LYMPHOCYTE	9.6	7.88	7.2	7.4										1	8
33. WBC COUNT	3.8	5.2	7.1	4	4.5									1	5
34. VIRAL LOAD	8640	####	####	###	###									150105	105146
35. BLOOD CHEMISTRY (SODIUM)						141	141	140	140	143	141	137		2	140
36. CREATININE						91	61	109	107	94	100	116		18	97
37. LIVER TEST (TOTAL PROTEIN)						82	92	85	86	95	87	94		5	89
38. LIVER TEST (ALT)														#DIV/0!	#DIV/0!

PART 4: RECORD OF VITAL OUTCOMES ..... Please; honestly and accurately complete each PURPLE box by using data from the patient file.

39. Date of last follow-up	6/30/2017	41. Vital status at Last Follow-up	
40. Total length of follow-up (months)	87	Transferred out$_1$ LTF$_2$ Dead$_3$ Alive at the end of follow-up$_4$	4

Appendix B: Research Data storage tool (Excel)

UNISA

CODED AND PROGRAMMED DATA TOOL

ANALYSIS OF THE SURVIVAL TIME OF HIV+ TERMINAL PATIENTS FROM ALBERT LUTHULI MUNICIPALITY IN RSA

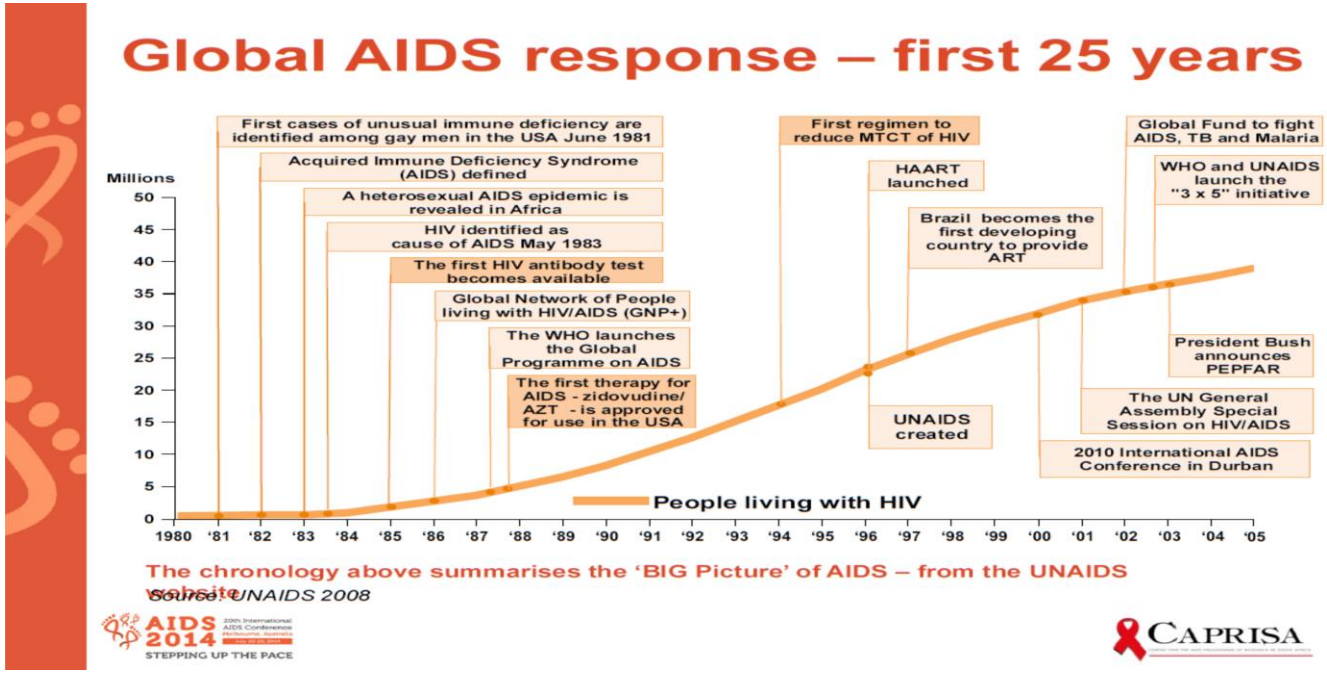
COLLECTED PATIENTS DATA FOR SURVIVAL ANALYSIS   EMBHULENI HOSPITAL   (MALES)

health MPUMALANGA PROVINCE REPUBLIC OF SOUTH AFRICA

ID	FOLDER NUMBER	GENDER 1	HOSP 2	HIVD 3	NATI 4	SEX 5	ALCO 6	WHOS 7	OPPI 8	RACE 9	MARI 10	FRHA 11	ARTR 1 12a	ARTR 2 12b	ARTR CODE 12c	ARTA 13	TBHI 14	DOB 15	ARTSD 16	LFD 39	BLMAS 17	BLCD4C 18	BLHAEM 19	BLLYMP 20	BLWBC 21	BLVIRL 22
1	09/14245	2	2	1	1	1	2	2	1	2	1	8	18	3	2A3L	1	2	7/28/76	3/3/10	6/30/17	72	114	13.6	10.9	3	40
2	10/1811	2	2	1	1	1	2	1	1	2	1	8	19		1TFE	2	2	9/2/78	3/10/10	6/30/17	68	207	14	30.5	3.4	49
3	13/2582	2	2	1	1	1	2	1	1	2	3	8	20		1TFE	2	2	2/4/77	3/1/10	6/30/17	68	230	11	36	4.2	15
4	09/2353	2	2	1	1	1	2	2	1	2	3	8	19		1T3N	2	2	7/15/69	2/1/10	6/30/17	82	189		14.1	5.8	175
5	07/2290	2	2	1	1	1	2	4	1	2	3	8	18		1TFE	2	1	2/3/74	1/11/10	6/30/17	68	60		7.9	8.9	75
6	98/9741	2	2	1	1	4	2	1	1	2	1	8	6	5	2Z3L	2	2	10/15/97	1/6/10	6/30/17	45	328	9.5	15.9	4.8	####
7	09/15597	2	2	1	1	1	2	1	1	2	2	8	18	17	1TFE	2	2	6/13/65	3/10/10	4/25/17	58	140	15	8.5	4.4	46
8	06/12372	2	2	1	1	1	2	1	4	2	2	7	18		1TFE	1	2	11/3/69	2/17/10	4/5/16	46	87	7.6	7.3	5.9	75
9	03/11647	2	2	1	1	1	2	3	1	2	1	0	5		1S3E	2	2	7/10/61	3/24/10	8/21/10	53	177	13	11.3	2.8	
10	10/986	2	2	1	1	4	2	3	1	2	1	0	5		1S3E	2	2	3/27/02	2/17/10	6/15/10	25	459	12	9.4	6.3	
11	10/3063	2	2	1	1	1	2	2	1	2	2	0	5		1S3E	2	2	12/4/67	3/31/10	6/19/10	64	53	9.6	5.1	3.3	
12	10/1152	2	2	1	1	4	2	2	1	2	1	0	1		1S3L	2	2	11/18/09	3/15/10	5/15/10	5	2647		30	15	####
13	09/7651	2	2	1	1	1	2	3	1	2	1	0	5		1S3E	2	2	9/10/88	3/31/10	8/20/10	55	48	13	2.5	7	
14	09/13682	2	2	1	1	1	2	3	6	2	1	0	5		1S3E	2	1	7/15/84	1/1/10	8/24/10	51	86		4.3	8.7	
15	08/12471	2	2	1	1	1	2	3	1	2	2	1	5		1S3E	3	2	9/28/57	3/31/10	7/7/11	58	183	12	11.8	5.9	
16	10/2258	2	2	1	1	1	2	3	1	2	2	1	1		1S3N	2	2	10/10/65	3/19/10	5/20/11	56	224	14	20.1	4.1	
17	09/6715	2	2	1	1	1	2	3	1	2	1	2	5	19	1T3E	1	2	9/17/83	3/19/10	11/19/11	57	246	8	29.1	2.3	####
18	09/13731	2	2	1	1	1	2	1	6	2	1	5	18	17	1TFE	2	2	11/11/68	1/4/10	8/24/14	59	164		22.5	4.6	55
19	07/4912	2	2	1	1	4	2	1	1	2	2	7	5	14	1A3E	2	2	3/27/01	3/31/10	2/24/16	37	411		21	5	50
20	08/4706	2	2	1	1	4	2	1	1	2	1	1	14	4	1A3E	2	2	5/7/02	1/18/10	1/17/11	30	219		21	5	658
21	09/10905	2	2	1	1	4	2	3	1	2	2	5	5	14	1A3E	2	2	5/25/04	1/5/10	7/11/14	27	330		37	6.6	220
22	06/6288	2	2	1	1	1	2	1	1	2	1	8	18	19	1TFN	2	2	9/13/76	3/31/10	6/30/17	56	77		27	6	94
23	10/361	2	2	1	1	1	2	3	1	2	2	8	18		1TFE	2	2	4/17/76	1/20/10	6/30/17	64	76		12.4	3.4	367
24	04/11764	2	2	1	1	1	2	4	6	2	2	8	18		1TFE	3	2	4/15/56	2/22/10	6/30/17	67	100	14	4.7	5	74
25	04/10824	2	2	1	1	1	2	1	1	2	1	6	19		1TFN	1	2	7/6/61	3/31/10	12/11/15	100	158		17	9.1	346
26	05/9411	2	2	1	1	1	2	3	1	2	1	8	5	8	1TFE	2	2	3/2/72	2/3/10	6/30/17	87	109	9.8	8.9	2.9	42
27	04/9831	2	2	1	1	4	2	3	1	2	1	8	5	13	1TFE	2	2	7/5/04	3/24/10	6/30/17	18	787	9.9	18	9.4	605
28	07/229	2	2	1	1	1	2	1	6	2	1	8	19	8	2Z3L	3	2	2/14/82	3/31/10	4/13/17	81	224	15	11.4	3.5	5800
29	00/9605	2	2	1	1	1	2	2	1	2	1	8	5	8	1TFE	2	2	12/23/56	3/30/10	6/30/17	50	61		5.6	3.8	124
30	09/15650	2	2	1	1	1	2	4	1	2	3	8	18	8	1TFE	3	2	3/20/77	2/10/10	6/30/17	64	239	8.5	7.6	7.2	94
31	06/9920	2	2	1	1	4	2	3	1	2	1	8	5	14	1A3E	3	2	4/17/98	2/22/10	6/30/17	25	102	12	6.6	12	####
32	10/1945	2	2	1	1	1	2	2	1	2	2	1	5		1S3E	2	2	6/14/69	3/24/10	8/24/11	56	75		12.6	6.8	66
33	10/5001	2	2	1	1	1	2	3	1	2	4	1	5		1S3E	2	2	10/10/73	3/26/10	8/24/11	56	44	11	4.6	6.3	124
34	09/12887	2	2	2	1	2	2	3	6	2	2	0	5		1S3E	2	1	1/11/74	2/26/10	12/7/10	48	20	10	5	5.6	
35	08/3514	2	2	1	1	1	2	3	6	2	1	0	5		1S3E	2	1	3/5/88	1/18/10	9/18/10	53	126	13	3	3.3	124
36	09/067	2	2	1	1	1	2	3	1	2	2	2	5		1T3E	2	1	4/30/74	1/4/10	6/28/12	50	47	9.8	10.9	2.9	20
37	09/14499	2	2	1	1	1	2	3	1	2	2	7	6		1Z3L	2	2	12/25/74	1/6/10	12/30/15	94	871	15.5		4.9	124
38	06/3859	2	2	1	1	4	2	3	1	2	1	0	5		1S3E	2	2	2/2/00	3/22/10	3/30/10	24	174	13	12.3	8.9	
39	99/12012	2	2	1	1	1	2	3	1	2	5	1	5		1S3E	3	2	12/23/73	3/15/10	10/15/11	61	77		31.4	6.1	124
40	10/658	2	2	1	1	1	2	3	1	2	1	0	5		1S3E	1	1	8/30/68	3/1/10	7/1/10	64	128	7.2	1.5		
41	09/10264	2	2	1	1	1	2	3	6	2	3	4	5		1S3E	2	1	7/26/77	3/16/10	9/16/10	66	173	7.2	3.3	5.1	
42	02/12724	2	2	1	1	1	2	3	1	2	2	1	18		1S3E	2	2	1/11/75	2/8/10	7/8/11	53	42	14	3.9	2.9	####
43	10/570	2	2	1	1	1	2	2	1	2	2	1	5		1S3E	2	2	6/30/73	2/11/10	7/11/11	62	187	15	17	2.7	124
44	09/2965	2	2	1	1	1	2	3	1	2	2	0	5		1TFE	2	2	12/12/54	1/4/10	6/19/10	60	170	14	16.4	6.1	
45	09/14337	2	2	1	1	1	2	3	6	2	2	7	5		2S3L	2	2	2/19/75	1/19/10	3/15/16	51	157	13	7.5	20	3380
46	09/14428	2	2	1	1	1	2	3	1	2	3	0	5		1S3E	1	2	12/15/73	1/6/10	3/6/10	57	131	17	8.4	11	124
47	10/847	2	2	1	1	1	2	3	6	2	2	0	5		1S3E	2	1	10/14/64	2/15/10	6/15/10	53	116	9.1	15.1	4.8	
48	04/9280	2	2	1	1	1	2	2	1	2	2	2	5		1T3E	2	2	6/24/64	3/9/10	4/9/12	92	218	11	11.2	7	124
49	09/15668	2	2	1	1	1	2	3	1	2	5	1	5		1S3E	2	2	1/1/53	1/13/10	6/9/11	93	34	12	2.3	2.9	124
50	09/15142	2	2	1	1	1	2	3	1	2	1	0	1		1S3N	1	2	3/9/66	1/18/10	3/18/10	60	12	11	1.5	2.7	
51	09/14168	2	2	1	1	1	2	3	1	2	2	0	5		1S3E	1	2	11/12/63	2/18/10	1/25/10	75	109	10	13.8	7.5	
52	09/14848	2	2	1	1	4	2	2	1	2	1	0	5		1S3L	2	2	9/25/09	2/8/10	5/8/10	5.5	600	10	21.2	3.4	####
53	09/14204	2	2	1	1	1	2	3	6	2	2	0	5		1S3E	2	1	4/17/57	1/11/10	8/11/10	81	244		23.7	5.8	
54	09/6800	2	2	1	1	1	2	4	6	2	3	0	1		1S3E	1	2	1/8/83	1/5/10	4/5/10	41	119	16	9.7	4.5	

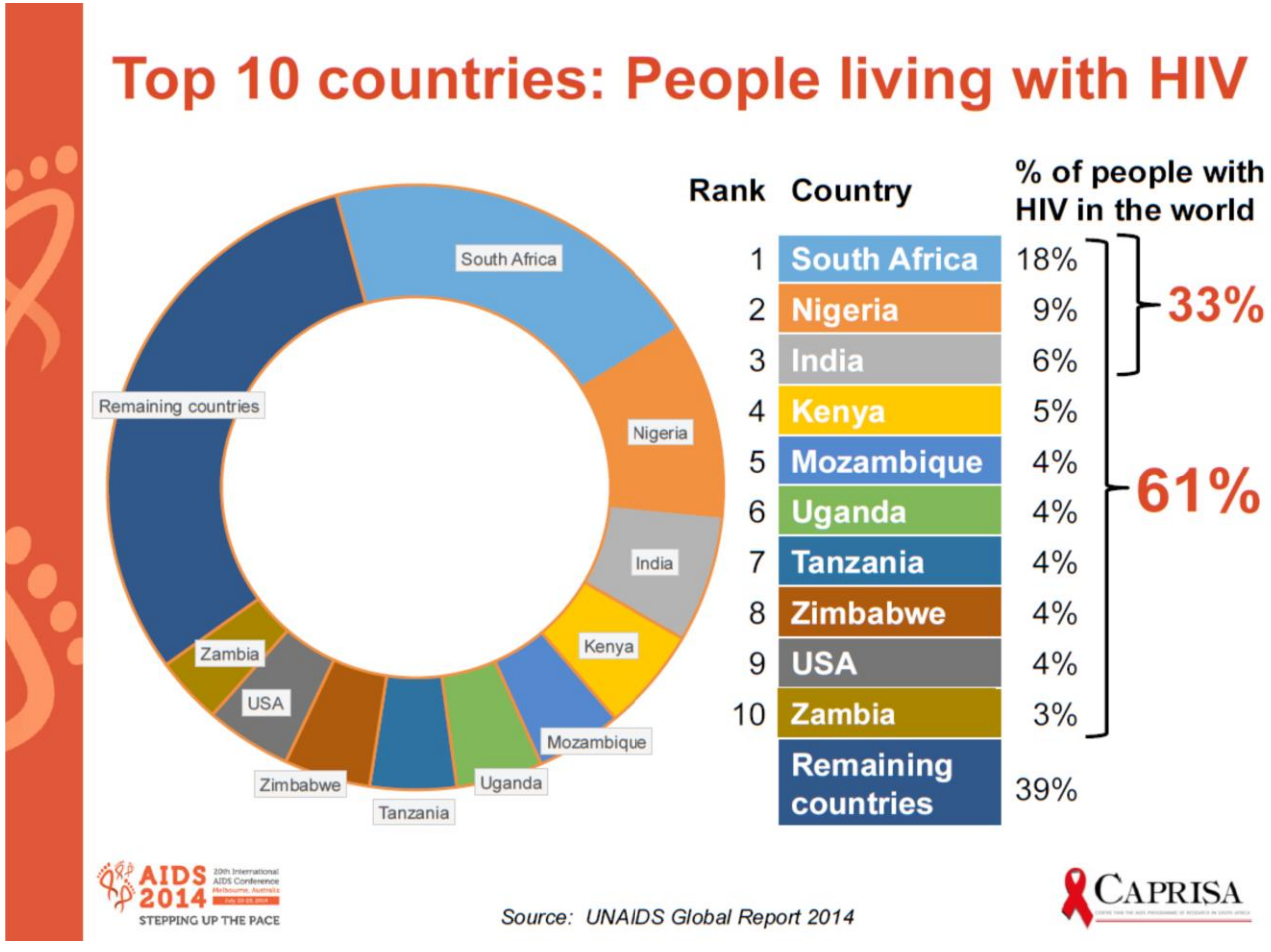
ID	AGE 23	BLCREAT 24	BLTPROT 25	BLSODIUM 26	FLKIDNEY 27	BLLIVER TEST 28	FLMAS 29 AV	FLMAS 29 SD	FLCD4 30 AV	FLCD4 30 SD	FLHAEM 31 AV	FLHAEM 31 SD	FLLYMP 32 AV	FLLYMP 32 SD	FLWBC 33 AV	FLWBC 33 SD	FLVIRL 34 AV	FLVIRL 34 SD	FLSODIUM 35 AV	FLSODIUM 35 SD	FLCREAT 36 AV	FLCREAT 36 SD	FLTPROT 37 AV	FLTPROT 37 SD	FLLIVER 38 AV	FLLIVER 38 SD	TLF 40	VSLF 41
1	33	116		137			83	5	131	74	15		8	1	5	1	####	####	140	2	97	18	89	5			87	4
2	31	76	77	140			74	6	328	123	14		29	6	4	1	207	223	142	2	98	19	82	6			87	4
3	32	140	69	137		13	69	1	714	73	11		35	2	5	1	15		137	2	169	22	78	6	20	6	87	4
4	40	104	79	143	26	91	85	3	437	122			29	4	4	1	2206	2554	143	2	93	19	84	11	96	87	88	4
5	35	32	96	148	28	16	68	2	431	130			18	6	8	1	565	668	142	3	80	23	89	9	18	5	89	4
6	12	43	99	141		10	49	3	268	181	10		10	5	8	3	####	####	140	3	49	7	92	13	10	5	89	4
7	44	80	99	141	26	22	55	3	242	102	15		18	9	7	2	184	150	144	10	85	15	90	9	36	18	85	4
8	40	76	91	140	27	10	45	1	325	143	7		19	8	5	1	45	42	141	3	73	16	88	8	15	7	73	1
9	48	117	80	139		52	57	3	177		13		11		3	1			139		117		80		52		4	1
10	7	42	99	135		18	24	1	461	3	12		11	2	7				135		42		98	1	21	4	4	1
11	42	57	116	140		29	67	3	53		10		5		3				140		57		116		29		2	1
12	0						5		###				30		15		####										2	1
13	21	85	76	140		62	59	4	48		13		3		5	2			140		85		76		62		4	1
14	25		100				61	7	224	195			11	10	7	3							100		66		7	1
15	52	134	104	136		35	60	1	284	143	12		14	3	6	2	82		136	1	105	42	112	11	23	18	15	1
16	44	98	84	140		15	59	3	321	136	14		23	5	5	1			141	1	98		84		27	17	14	1
17	26	54	92	141		12	55	2	266	37	8		29		4	3	####		140	2	54		94	3	13	1	20	3
18	41	89	71	138	28	31	59	2	543	202			31	5	5	1	35	28	139	2	101	11	73	3	33	12	7	1
19	8	45	89	141		15	43	5	702	196			26	4	8	2	747	1515	141	2	62	28	92	15	21	6	70	2
20	7	60	75	140		30	37	5	364	167			21	9	7	2	####	####	141	2	46	20	79	7	29	9	11	2
21	5	33	77	140		16	36	7	542	142			35	3	6	2	673	1009	142	2	38	5	83	8	21	8	54	1
22	33	55	79	140	24	11	58	2	684	108			33	6	5	1	673	1015	142	3	89	18	76	6	28	15	86	4
23	33	86	93	140	26	21	67	2	264	63			18	4	4	1	190	161	142	2	82	14	82	8	24	7	89	4
24	53	101	81	135	27	65	65	2	268	104	14		18	7	5	1	571	931	139	3	92	9	75	6	37	16	88	4
25	48	77	81	140	25	58	99	1	462	122			21	3	9	1	1025	1722	142	3	97	18	78	5	42	13	68	1
26	37	80	120	135		58	103	7	379	159	10		18	5	6	1	141	130	140	4	83	16	97	14	45	16	88	4
27	5	65	118	138		23	29	9	###	205	10		28	5	9	1	258	301	140	2	53	16	89	15	26	12	86	4
28	27	73	87	139	25	34	83	7	171	121	15		8	5	4	1	####	####	141	2	88	13	85	11	47	21	84	1
29	53	55	83	157	25	32	50	1	###	527			36	18	6	1	124		144	9	74	16	83	3	23	8	87	4
30	32	196	81	132		52	62	4	540	178	7	2	21	10	8	2	56	35	140	4	105	40	81	10	29	14	88	4
31	11	81	109	143		46	31	4	159	52	12		8	3	8	3	####	####	139	3	58	23	94	14	18	14	88	4
32	40	73	68	142		10	56	1	246	133			14	2	7	1	95	41	142		73		68		10		17	1
33	36	91	86	130		64	56	1	171	127	11		8	5	6		124		135	7	82	13	96	14	59	8	16	1
34	35		69	134		11	53	4	42	30	10		5		6				134				69		11		9	1
35	21	77	95	142		72	54	1	311		13		3		3		124		142		77		95		72		8	1
36	35	80		135		27	61	6	150	104	10		11		3		20		135		80				27		29	2
37	35	73	71	143		24	94		890		16				5		124		143		73		71		24		71	1
38	9						25	1	493	283	13		15	4	12	5											0	1
39	36	38	116	138		505	59	2	416	121			33	2	7	1	124		138		38		116		505		19	1
40	41						60	4	128		7		2														4	1
41	32	59	95	135			71	7	117	80	7		3		5				135		59		85	15	36		6	1
42	34	64	109	141		37	61	5	50	7	14		3	1	3	1	####		141		64		109		37		17	1
43	36	80	82	138		19	63	1	640	640	15		17		3	1	124		138		80		82		19		17	1
44	55		78	141		54	61	1	254	119	14		16		7	2			141				76	4	55	1	5	1
45	34	97	110	122	1	276	56	2	298	177	13		17	6	6	6	####	####	135	9	90	11	83	26	11	143	73	1
46	36		98	123		32	57		210	111	17		8		11		124		123				98		32		2	2
47	45	117	76	139		10	56	3	116		9		15		5				139		117		76		10		4	2
48	45	67	89	137		26	99	5	346	181	11		11		8	1	124		140	4	70	4	93	5	32	8	25	2
49	57	74	70	138		46	104	7	243	211	12		4	3	5	2	124		138	1	74		71	1	69	32	16	1
50	43	74	88	139		26	60		12		11		2		3				139		74		88		26		2	1
51	46	88	75	124		36	75		109		9	2	14		9	1			124		88		75		36		0	2
52	0						6	1	259		10		3		####												3	2
53	52						85	5	211	57			25	2	9	5											7	2
54	26	78	79	134	1	374	41		119		16		10		6	3			134		78		79		374		3	2

APPENDIX C: Summary of the big picture of AIDS



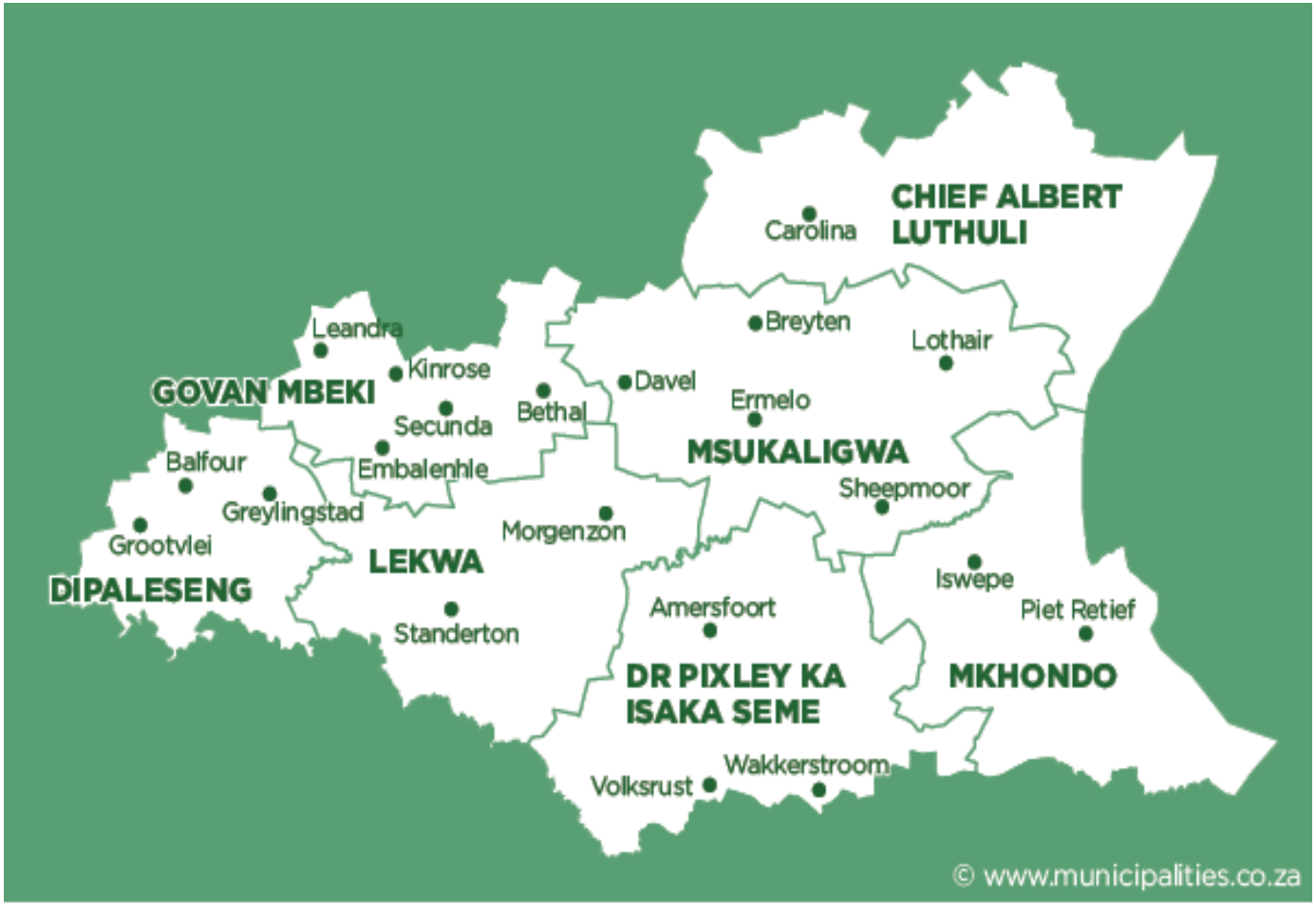
Source: Karim S.S.A. (2014). State of the Art: Epidemiology and Access. *UNAIDS, CAPRISA*.

APPENDIX D: South Africa in first position of people living with HIV



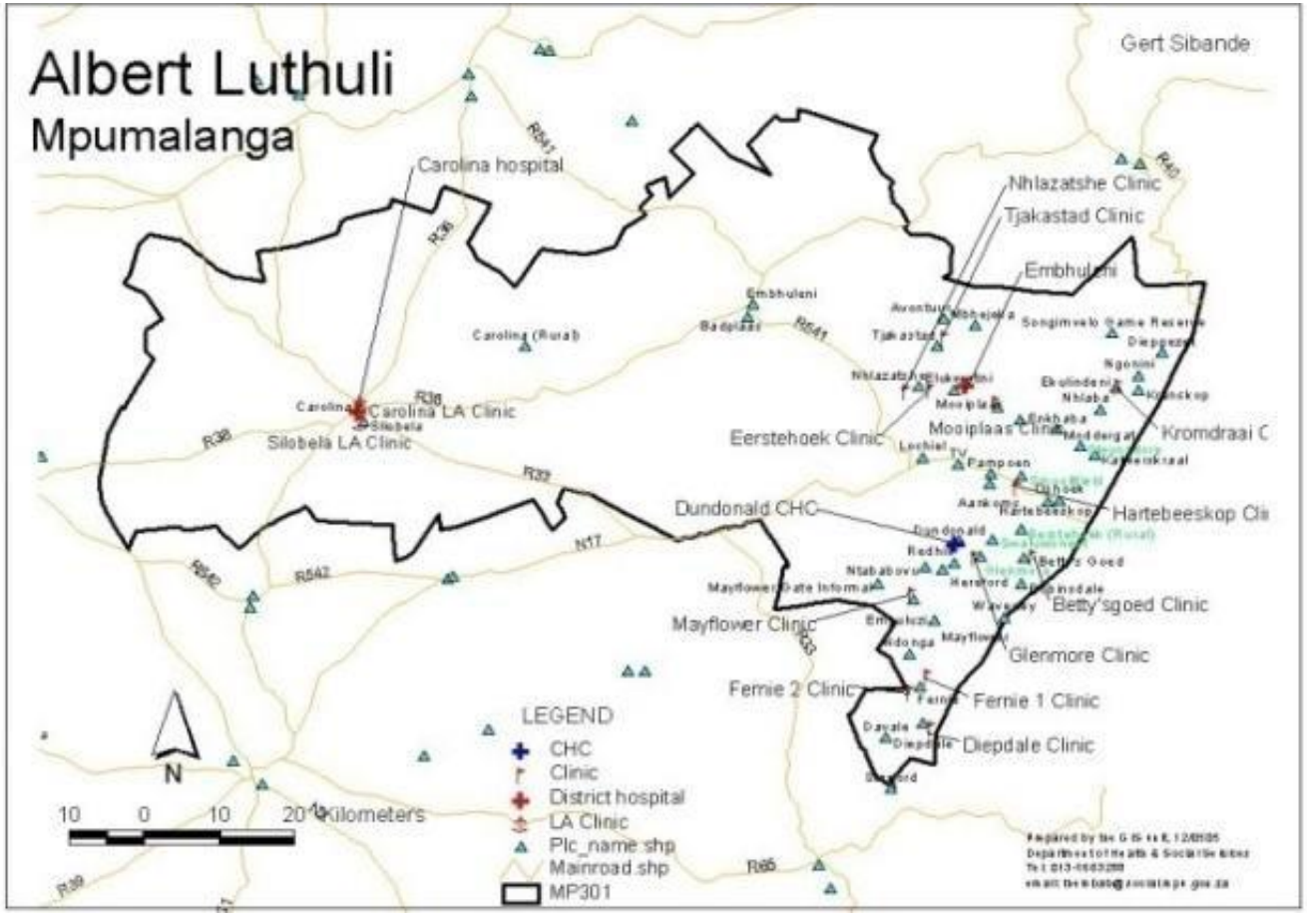
Source: Karim S.S.A. (2014). The HIV Epidemic: Progress & Challenges Southern African HIV Clinicians Society Conference, *UNAIDS Global Report 2014.*

Appendix E: Map showing the relative position of Chief Albert Luthuli in Gert Sibande District



Source: Mpumalanga Municipalities. (2017). Chief Albert Luthuli Local Municipality (MP301).

Appendix F: Map showing the distribution of the Health sites in Albert Luthuli Municipality



Source: Ogunsanwo, D.A (2012). Determination of patient satisfaction at accredited antiretroviral treatment sites in the Gert Sibande District, Mpumalanga Province.

Appendix G: Important statistics on Albert Luthuli Municipality.

Population	2016	2011
	187 629	186 010
Age Structure		
Population under 15	34.1%	36.5%
Population 15 to 64	60.3%	58.2%
Population over 65	5.7%	5.3%
Dependency Ratio		
Per 100 (15-64)	65.9	71.7
Sex Ratio		
Males per 100 females	89.1	88.2
Population Growth		
Per annum	0.20%	n/a
Labour Market		
Unemployment rate (official)	n/a	35.4%
Youth unemployment rate (official) 15-34	n/a	45.1%
Education (aged 20 +)		
No schooling	15.5%	19.9%
Matric	31.8%	27.0%
Higher education	5.8%	6.3%
Household Dynamics		
Households	53 480	47 705
Average household size	3.5	3.8
Female-headed households	48.4%	49.3%
Formal dwellings	80.2%	76.5%

Source: Mpumalanga Municipalities. (2017). Chief Albert Luthuli Local Municipality (MP301). Statistics South Africa (Stats SA) (2011). Census 2011 Municipal report.

Appendix H: Descriptions of technical independent variables part 1

Variable	Definition	Notes
CD4	CD4 cell count - is the measurement of the number of blood cells in a cubic millimeter of blood. **CD4-** A molecule on the surface of some cells onto which HIV can bind.CD4 cell percentage-proportion of all white blood cells that are CD4 cells.	The CD4 cell count of a person not infected with HIV can be between 500 and 1500.A CD4 cell count between 200 and 500 indicates that some damage to the immune system has occurred. It is particularly important to have CD4 cell count carefully monitored as it gets closer to 350. If CD4 count is below 200 the HIV disease is classified as AIDS.
Viral load	The number of copies of HIV RNA in a millilitre of blood. Viral load describes the amount of HIV in the blood.	The more HIV there is in the blood then the faster the CD4 cell count will fall, and the greater the risk of becoming ill because of HIV. Amongst people with the same CD4 cell count, those with a high viral load tend to lose CD4 cells and become ill faster. If the viral load is below 50, it is usually said to be undetectable. The aim of HIV treatment is to reach an undetectable viral load. High viral loads are linked to faster disease progression.
White blood cell count	This is a measure of the total number of white blood cells	These cells are part of the immune system and defend against infections. People with HIV often have slightly lower levels. High white blood cell counts may indicate the body is fighting an infection. Low counts put the body at risk of getting an infection.
Total protein	It measures the total amount of albumin and globulin in the body.	It may be used in the event of unexpected weight loss, fatigue or symptoms of kidney or liver disease.

Main source: Corkery, S. (2016). Diagnosed with HIV at a low CD4 count. *Nam Aidsmap HIV&AIDS sharing knowledge, changing lives.*

Appendix I: Descriptions of technical independent variables part 2

Variable	Definition	Notes
Creatinine and Sodium (Electrolyte test)	Creatinine and Sodium are chemical wastes generated from muscle metabolism and their levels help measure how well the kidneys are working.	The kidneys filter out most of the creatinine and sodium and dispose of them in the urine. The kidneys maintain the blood creatinine and sodium levels in a normal range. Elevated creatinine and sodium levels signify impaired kidney function or kidney disease.
Alanine aminotransferase [ALT] (Liver functions test)	Liver disease is often reflected by biochemical abnormalities of liver function. Elevated serum activity of the two commonly used liver enzymes [ALT] and [AST]) reflects liver cell injury.	Many authors agree that elevated serum activity of the two commonly used liver enzymes (alanine aminotransferase [ALT] and aspartate aminotransferase [AST]) that are involved in breakdown of amino acids reflects liver cell injury.
Total lymphocyte counts (TLC)	Lymphatic system is made up of tissues and organs like spleen, tonsils and lymph nodes which protect the body from infection.	Studies have shown good correlation between total lymphocyte counts (TLC) and CD4 count. TLC has the advantage of being less expensive and less complicated than the CD4 count in HIV-related diagnosis
Haemoglobin	Red blood cells of the blood which are essential for transferring oxygen in the blood from the lungs to the tissues.	As the HIV infection progresses, blood haemoglobin decreases. Haemoglobin levels could be measured easily where resources for measuring CD4 are limited and this could help to determine patients who are at greatest risk of disease progression, allowing these patients to be identified for closer monitoring or therapeutic intervention.

Main source: Corkery, S. (2016). Diagnosed with HIV at a low CD4 count. *Nam Aidsmap HIV&AIDS sharing knowledge, changing lives.*

Appendix J: Laboratory Report on Sodium, Creatinine, Kidney damage (eGFR) Total protein and ALT

NATIONAL HEALTH LABORATORY SERVICE

EMBHULENI LABORATORY
Embhuleni Hospital, Stand No 40b, Diepgesig Road, Elukwatini
Mpumalanga, 1192
Tel 017 883 1504, Fax 017 883 2560
pg 1 of 2

Practice Number

FULL FINAL LABORATORY REPORT

PATIENT:

LAB NUMBER: PD

REPORT TO:

31/12/1980 (31y) Sex F

Ref Number:
Collected: 12/05/2012 ?
Received: 12/05/2012 14:06
Printed: 14/05/2012 15:16

ART Clinic
Embhuleni Hospital
Private Bag X1001
Elukwatini
Mpumalanga
1192

Patient Location: Embhuleni Hospital, ART Clinic
Hospital Number:

FOR ENQUIRIES AND FOLLOW-UP TESTS, PLEASE QUOTE PATIENT'S MRN NUMBER

CHEMICAL PATHOLOGY

Specimen received: Blood
Tests requested: Na, K, Cl, Urea, Creat, TP, Alb, T bili, C bili, ALT, AST, ALP, GGT

Blood chemistry:			
Sodium	146 H	mmol/L	136 - 145
Potassium	Innumerable		
Chloride	106	mmol/L	98 - 107
Urea	4.1	mmol/L	2.1 - 7.1
Creatinine and estimated clearance:			
Creatinine	16 L	umol/L	49 - 90
eGFR (MDRD formula)	>60	mL/min/1.73 m^2	

Although GFR estimation using the MDRD (4-variable) equation is recommended for patients with chronic kidney disease (CKD) and risk factors for CKD, it significantly underestimates true GFR in patients with normal renal function. The eGFR reported here has not been adjusted by any factor for race.

Estimates of GFR by the MDRD equation may be unreliable in: patients younger than 18 years or older than 70 years of age; pregnancy; the presence of serious morbid conditions; acute renal failure; extremes of body size/nutritional status/muscle mass or unusual dietary intake.

Staging of Chronic Kidney Disease (KDOQI of National Kidney Foundation):

Stage	Description	GFR ml/min/1.73m^2
1 & 2	Kidney damage with mildly decreased or normal/increased GFR	> 60
3	Moderate decrease in GFR	30 - 59
4	Severe decrease in GFR	15 - 29
5	Kidney failure	< 15 (for dialysis)

Blood chemistry:			
Total protein	84 H	g/L	60 - 78
Albumin	36	g/L	35 - 52
Total bilirubin	3 L	umol/L	5 - 21
Conjugated bilirubin (DBil)	1	umol/L	0 - 3
Alanine transaminase (ALT)	38 H	U/L	7 - 35
Aspartate transaminase (AST)	36 H	U/L	13 - 35
Alkaline phosphatase (ALP)	123 H	U/L	42 - 98

Appendix K: Laboratory Report on Haemoglobin

NATIONAL HEALTH LABORATORY SERVICE

Temporary details
PO Box 1038, Johannesburg, 2000, South Afric
Cnr Hospital & de Korte Str., Johannesbur
Tel: (27 11) 489-9000 Fax: (27 11) 489-900

EMBHULENI

FAX->017,883,0044 M/EERSE/0

Labno :
Patient :
Gender : Female Age: 28 yrs 11 mths DoB: 31/12/1980

SR LUSENGA
EERSTEHOEK CLINIC
PRIVATE BAG X100
ELUKWATINI
1192

Ref Dr :
Clinic : EERSTEHOEK CLINIC
Taken : 17/11/09 10:44 Registered: 18/11/09 15:42
Reported : 18/11/09 16:34

LABORATORY REPORT

-- page 1 --

CLINICAL DATA : ?RVD
SPECIMEN : Blood
TESTS ORDERED: FBCPLT-CEL

FULL BLOOD COUNT & PLATELETS			Flags	Reference Ranges
White Cell Count	4.70	x 10^9/l		4.00 - 10.0
Red Cell Count	3.66	x 10^12/l	L	4.13 - 5.67
Haemoglobin	10.4	g/dl	L	12.1 - 16.3
Haematocrit	0.316	l/l	L	0.370 - 0.490
MCV	86.0	fl		79.1 - 98.9
MCH	28.3	pg		27.0 - 32.0
MCHC	32.8	g/dl		32.0 - 36.0
Red Cell Distribution Width ..	14.8	%	H	11.6 - 14.0
Platelets	293	x 10^9/l		178 - 400

Authorised by : AC Nkambule Medical Technologist

Prof HF Jowert

***** End of Laboratory Report *****

Total pages : 1

Appendix L: Laboratory Report on Viral Load

NATIONAL HEALTH LABORATORY SERVICE

Embhuleni Laboratory
Embhuleni Hospital, Stand No 40b, Diepgesig Road, Elukwatini, Mpumalanga, 1192
Tel 017 883 1504, Fax 017 883 2560

Pr No _____ pg 1 of 1

FULL FINAL LABORATORY REPORT

PATIENT:

LAB NUMBER: PD

REPORT TO:

31/12/1980 (31y) Sex F

Ref Number:
Collected: 10/05/2012 ?
Received: 11/05/2012 20:26
Printed: 13/05/2012 10:00

ART Clinic
Embhuleni Hospital
Private Bag X1001
Elukwatini
Mpumalanga
1192

Patient Location: Embhuleni Hospital, ART Clinic
Hospital Number:

FOR ENQUIRIES AND FOLLOW-UP TESTS, PLEASE QUOTE PATIENT'S MRN NUMBER

DEPARTMENT OF VIROLOGY

Specimen received: Blood
Tests requested: HIV Viral Load@
@ Test(s) performed in another laboratory

HIV Molecular Investigations:

HIV Viral Load:

HIV Viral Load	29560	copies/mL
Log Conversion	4.47	
Input Volume	1.000	mL
Methodology	Roche COBAS AmpliPrep/TaqMan HIV-1 Test, v2	

Comment:
Based on the sample volume tested, the lower limit of detection is 20 copies/mL. The linear range of this assay is 20 - 10,000,000 copies/mL (1.3 - 7 log copies/mL).

The result should be interpreted in conjunction with CD4 counts and the patient's clinical status. A result of lower than detectable limit cannot be presumed to be negative for HIV-1 RNA.

@ HIV Viral Load performed at Nelspruit Laboratory

Authorised by: IM Mofokeng (Medical Technologist) HIV Viral Load

-- End of Laboratory Report --

Appendix M: Laboratory Report on CD4 count, Lymphocyte and White blood cell count

NATIONAL HEALTH LABORATORY SERVICE

Embhuleni Laboratory
Embhuleni Hospital, Stand No 40b, Diepgesig Road, Elukwatini, Mpumalanga, 1192
Pr No 5200296 Tel 017 883 1504, Fax 017 883 2560 pg 1 of 1

FULL FINAL LABORATORY REPORT

PATIENT: LAB NUMBER: PD REPORT TO:

31/12/1980 (31y) Sex F

Ref Number:
Collected: 10/05/2012 ?
Received: 11/05/2012 20:20
Printed: 12/05/2012 19:45

ART Clinic
Embhuleni Hospital
Private Bag X1001
Elukwatini
Mpumalanga
1192

Patient Location: Embhuleni Hospital, ART Clinic
Hospital Number:

FOR ENQUIRIES AND FOLLOW-UP TESTS, PLEASE QUOTE PATIENT'S MRN NUMBER

DEPARTMENT OF HAEMATOLOGY

Specimen received: Blood
Tests requested: CD4@
@ Test(s) performed in another laboratory

CDARV:

CD45 +ve White Cell Count	4.28	x 10^9/L	4.00 - 10.00
CD4% of Lymphocytes	29.72	%	28.00 - 58.00
Absolute CD4	567.40	x 10^6/L	500.00 - 2010.00

@ CD4 performed at Nelspruit Laboratory

Authorised by: IM Mofokeng (Medical Technologist) CD4

-- End of Laboratory Report --

Appendix N: Mpumalanga Department of Health Permission letter

health
MPUMALANGA PROVINCE
REPUBLIC OF SOUTH AFRICA

No.3, Government Boulevard, Riverside Park, Ext. 2, Mbombela, 1200, Mpumalanga Province
Private Bag X11285, Mbombela, 1200, Mpumalanga Province
Tel I: +27 (13) 766 3429, Fax: +27 (13) 766 3458

Litiko Letemphilo | Departement van Gesondheid | UmNyango WezeMaphilo

Enquiries: Research (013) 766 3511/3757/3766

PEPUKAI BENGURA
UNISA - University Of South Africa

Dear Mr Bengura

APPLICATION FOR RESEARCH APPROVAL: IDENTIFICATION AND MODELLING OF FACTORS AFFECTING THE SURVIVAL LIFE TIME OF HIV+ TERMINAL PATIENTS IN ALBERT LUTHULI MUNICIPALITY IN SOUTH AFRICA

The Provincial Health Research Committee has approved your research proposal in the latest format that you sent.

- ***PHREC REF: MP_201708_013***
- ***Approval Period: June 2017 to June 2018***
- ***Approved Facilities: Carolina and Embhuleni Hospitals***
- ***Resources Requested: Nurses (4), Doctors (4) & Patient Files***

Kindly ensure that the study is conducted with minimal disruption and impact on our staff, and also ensure that you provide us with the soft and hard copies of the report once your research project has been completed.

Kind regards

MR J SIGUDLA
MPUMALANGA PHRC

MPUMALANGA PROVINCE
DEPARTMENT OF HEALTH
PLANNING & INFORMATION
2017 -09- 05
DATE
PRIVATE BAG X 11285
NELSPRUIT 1200

MPUMALANGA
THE PLACE OF THE RISING SUN

Appendix O: UNISA Ethics Approval

UNISA university of south africa

SCHOOL OF SCIENCE RESEARCH ETHICS REVIEW COMMITTEE

01 July 2017

Ref #: 2017/SSR-ERC/005
Name of applicant (student): Mr Pepukai Bengura
Student #: 31659322

Dear Mr Pepukai Bengura,

Decision: Ethics Approval

Name: Mr Pepukai Bengura, PO Box 26, Burgersfort, 1150, bpenaldo@yahoo.com, 0834535013

Supervisor: Prof P Ndlovu, Department of Statistics, 0116709250, ndlovp@unisa.ac.za

Co-Supervisor: MA Managa, Department of Statistics, 0116709000, managma@unisa.ac.za

Proposal: Identification and modelling of the factors affecting the survivial time of HIV+ terminal patients in Albert Luthuli Municipality in South Africa

Qualification: Postgraduate degree

Thank you for the application for research ethics clearance by the School of Science Research Ethics Review Committee for the above mentioned research. Final approval is granted for the duration of the project.

For full approval: *The application was reviewed in compliance with the Unisa Policy on Research Ethics by the School of Science RERC on 28 June 2017.*

The proposed research may now commence with the proviso that:

1) *The researcher/s will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.*
2) *Any adverse circumstance arising in the undertaking of the research project that is relevant to the ethicality of the study, as well as changes in the methodology, should be communicated in writing to the (Name of unit/sub-unit) Ethics Review Committee. An amended application could be requested if there are substantial changes from the existing proposal, especially if those changes affect any of the study-related risks for the research participants.*
3) *The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study.*

Note:
The reference number [top right corner of this communiqué] should be clearly indicated on all forms of communication [e.g. Webmail, E-mail messages, letters] with the intended research participants, as well as with the School of Science RERC.

Kind regards,

Prof SJ Johnston
Chair: School of Science Research Committee
Email: johnssj@unisa.ac.za
Tel: 011 670 9146

Prof I Naidoo
Director: School of Science

Appendix P: Main statistical programming used: SAS, R and Stata

SAS Proc codes

MACROS FOR SURVIVAL GRAPHS FOR COX REGRESSION, KAPLAN-MEIER AND LOGISTIC REGRESSION

```
ods trace on;
PROC LIFETEST DATA=MON PLOTS=(S) ;
TIME LENFOLDAYS *DIED(0);
STRATA GENDER;
RUN;

proc template;
source Stat.Lifetest.Graphics.ProductLimitSurvival;
run;

options ls=95;
data _null_;
infile 'http://support.sas.com/documentation/onlinedoc/stat/ex_code/121/templft2.html'
device=url;
retain pre 0;
input;
if index(_infile_, '</pre>') then pre = 0;
if pre then put _infile_;
if index(_infile_, '<pre>') then pre = 1;
run;


%macro SurvivalTemplateRestore;
%global TitleText0 TitleText1 TitleText2 yOptions xOptions tips
tipl groups bandopts gridopts blockopts censored censorstr;
%let TitleText0 = METHOD " Survival Estimate";
%let TitleText1 = &titletext0 " for " STRATUMID;
%let TitleText2 = &titletext0 "s"; /* plural: Survival Estimates */
%let yOptions = label="Survival Probability" shortlabel="Survival"
linearopts=(viewmin=0 viewmax=1
tickvaluelist=(0 .2 .4 .6 .8 1.0));
%let xOptions = shortlabel=XNAME offsetmin=.05
linearopts=(viewmax=MAXTIME tickvaluelist=XTICKVALS
tickvaluefitpolicy=XTICKVALFITPOL);
. . .
%macro SurvivalTemplate; . . . %mend;
%macro entry_p; . . . %mend;
%macro SingleStratum; . . . %mend;
%macro MultipleStrata; . . . %mend;
%macro AtRiskLatticeStart; . . . %mend;
%macro AtRiskLatticeEnd; . . . %mend;
%SurvivalTemplate
%mend;


data _null_;
```

```
infile 'http://support.sas.com/documentation/onlinedoc/stat/ex_code/121/templft2.html'
device=url;
file 'junk.junk';
retain pre 0;
input;
if index(_infile_, '</pre>') then pre = 0;
if pre then put _infile_;
if index(_infile_, '<pre>') then pre = 1;
run;

%inc 'junk.junk' / nosource;

%SurvivalTemplateRestore
%SurvivalTemplate

proc template;
delete Stat.Lifetest.Graphics.ProductLimitSurvival / store=sasuser.templat;
delete Stat.Lifetest.Graphics.ProductLimitSurvival2 / store=sasuser.templat;
run;

%SurvivalTemplateRestore /* variables available */
%let TitleText0 = "Kaplan-Meier Plot"; /* Change the title. */
%let TitleText1 = &titletext0 " for " STRATUMID;
%let TitleText2 = &titletext0;
%SurvivalTemplate /* Compile the templates with */
/* the new title. */
proc lifetest data=MON /* Perform the analysis and make */
plots=survival(cb=hw test); /* the graph. */
time LENFOLDAYS * DIED(0);
strata GENDER;
format gender monfmt;
run;
%SurvivalTemplateRestore /* Restore the default macros */
/* and macro variables. */
proc template; /* Restore the default templates. */
delete Stat.Lifetest.Graphics.ProductLimitSurvival / store=sasuser.templat;
delete Stat.Lifetest.Graphics.ProductLimitSurvival2 / store=sasuser.templat;
run;
%SurvivalTemplateRestore
%let yOptions = label="Survival Probability"
linearopts=(viewmin=0.5 viewmax=1
tickvaluelist=(0.5  .6  .7 .8  .9   1.0));
%let xOptions = label="Follow-up time (days) " offsetmin=.05
linearopts=(viewmax=MAXTIME tickvaluelist=XTICKVALS
tickvaluefitpolicy=XTICKVALFITPOL);

%SurvivalTemplate
proc lifetest data=MON plots=survival( test);
time LENFOLDAYS * DIED(0);
strata GENDER;
format gender monfmt;
run;

data BEN;
set MON;
```



```
if baselineSODIUM>145 then baseline_sodium='Above normal';
else if baselineSODIUM<145 and baselineSODIUM >136 then baseline_sodium ='Normal';
else baseline_sodium ='Below normal';
run;

%modstyle(name=mystyle, parent=htmlblue,
colors=green red blue, fillcolors=green red blue)
ods html style=htmlbluecml image_dpi=300;

proc lifetest data=BEN plots=survival (TEST);
time LENFOLDAYS * DIED(0);
strata baseline_sodium TREAT1 ;
run;
ods html close;
```

LINEARITY AND PH TESTS

```
PROC PHREG DATA = WORK.MON;
MODEL LENFOLDAYS*DIED(0) =
    TREAT1 TBHISTORY MARITALSTATUS HOSPITAL GENDER FOLLOWUPCD4
    FOLLOWUPSODIUM BASELINEMASS FOLLOWUPMASS
    BASELINELYMPHOCYTE FOLLOWUPHAEMOGLOBIN
    BASELINEHAEMOGLOBIN FOLLOWUPVIRALLOAD ;

ASSESS VAR = (
    FOLLOWUPCD4 FOLLOWUPSODIUM BASELINEMASS FOLLOWUPMASS
    BASELINELYMPHOCYTE FOLLOWUPHAEMOGLOBIN
    BASELINEHAEMOGLOBIN FOLLOWUPVIRALLOAD)

PH / RESAMPLE SEED = 12345;
RUN;
```

R Commands

Sample size determination

MyData<-read.csv(file="C:/Users/PEPUKAINELSON/Desktop/RESEARCH DATA/FINAL/model2019 R.csv", header = TRUE, sep = ",")
attach(MyData)

```
hr=7.7
delta=0.15
pE=0.115
pA=0.115
alpha=0.05
beta=0.20
(n=((qnorm(1-alpha)+qnorm(1-beta/2))/(delta-abs(log(hr))))^2/(pA*(1-
pA)*pE))
ceiling(n)
```

QUANTILE REGRESSION

MyData<-read.csv(file="C:/Users/PEPUKAINELSON/Desktop/RESEARCH DATA/FINAL/model2019DECIDE.csv", header = TRUE, sep = ",")
attach(MyData)
x<-Age
y<-LENFOLdays
plot(x,y)
require(quantreg)
plot(x, y,cex=.25,type="n",xlab="Age",
ylab="Lenfoldays")
points(x, y,cex=.5,col="green")
abline(rq(y~x,tau=.5),col="yellow")
abline(lm(y~x),lty=2,col="red")
taus <- c(.05,.1,.5,.75,.95)
f <- rq(y~x, tau = taus)
for(i in 1:length(taus)){abline(coef(f)[,i],col="gray")}

SURVIVAL CURVES

plot(survfit(Surv(LENFOLdays,DIED)~Treat1))
plot(survfit(Surv(LENFOLdays,DIED)~Treat1),main="Plot of T",xlab="Time",ylab="Survival proba",col=c("blue","red","green","pink"))
legend("bottomright",
legend=c("NVP","EFV","TDF","TDF"),fill=c("blue","red","green","pink"),bty="n")

COX MODELLING

attach(MyData)

coxph(Surv(LENFOLdays,DIED)~FollowUpCD4+Age+Treat1+Hospital+WHOStage+FollowUpALT+ARTAdherence*Hospital+FollowUpSodium+ARTAdherence+MaritalStatus+BaselineViralLoad+

BaselineMass+FollowUpMass+FollowUpHaemoglobin+FollowUpLymphocyte+TBHistory+TBHistory*Hospital+BaselineViralLoad*WHOStage+ BaselineViralLoad*Treat1+Gender)

Stata commands

If you want to…	**then use command….**	**useful options (things you put after the comma)**
Tell STATA you have a censored outcome with observation time *obstime* and event indicator *status*	**stset** followuptime, died(0)	some exist in STATA, but you won't need them currently
Graph a Kaplan-Meier Survival curve	**sts graph by**(Gender)	allows you to graph separate survival curves on same plot by level of *varname* **gwood** adds Greenwood 95% CIs around the survival curve
Graph a smoothed curve of the baseline hazard function using kernel density estimation (based on the data, not any model)	**sts graph**	**hazard** makes a graph of baseline hazard instead of K-M survival curve **cihazard** adds confidence bands around hazard function
List Kaplan-Meier survival estimates	**sts list by**(TBhistory)	lists K-M survival estimates by levels of *varname*
Perform a logrank test of equality of survival functions	**sts test** treat1	**wilcoxon** allows you to do a Wilcoxon-Breslow test instead of a logrank test
Fit a Cox Proportional Hazards model **stcox** *predictor list*	**stcox** *predictor list*	**nohr** gives estimated beta coefficients instead of hazard ratios **robust** provides robust standard error estimates **cluster(***id***)** lets STATA know you have dependent data by id **schoenfeld(***varname*)* stores Schoenfeld residuals for each observation in a variable for each predictor in *predictor list* **scaledsch**(*varname**) stores scaled Schoenfeld residuals **mgale**(*varname*) stores Martingale

		residuals for each observation **basesurv**(*varname*) stores estimates of baseline survival in *varname* **basehc**(*varname*) stores estimates of baseline hazard in *varname*
Test the Proportional Hazards assumption (after fitting a Cox model)	**stphtest**	global test requires having specified schoenfeld() in stcox individual tests require having specified scaledsch() in stcox
Plot Scaled Schoenfeld residuals vs. time and look for flatness of the smooth (flat smooth implies PH is okay for that covariate, increasing or decreasing smooth implies problems with PH)	**lowess age**_t (here, *varname* is the stored scaled Schoenfeld residuals corresponding to covariate you are checking for PH)	**bwidth**(#) allows bandwidth specification by a number # (lower bandwidth makes curve try and fit each data point more)
Plot Martingale residuals for continuous covariates and look for shape of smooth (flat smooth implies functional form for covariate is okay, curved smooth implies transforming covariate in the model)	**lowess** followupcd4 *covariate* (here, *varname* is the stored Martingale residuals, *covariate* is a continuous covariate)	**bwidth**(#) allows bandwidth specification by a number # **mean** uses running-mean smoothing instead of least-squares smoothing(might want to do this when plotting Martingale residuals)