SURVIVAL ANALYSIS OF PROGNOSTIC FACTORS ASSOCIATED WITH CERVICAL CANCER

BY

SAYUTI, Fatima Yahaya
MTech/SPS/2019/10574

DEPARTMENT OF STATISTICS
FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA

JUNE, 2023

ABSTRACT

Cervical cancer is a major public health menace to women of reproductive age. It is the most common cancer in women in Nigeria. Survival refers to the life of a person after diagnosis of a disease and survival studies measure the overall performance of patients in terms of quality and quantity of life after diagnosis and treatment. This study investigate the data of patients diagnosed with cervical cancer from January 2010 to December 2020 at National Hospital Abuja (NHA). The Kaplan-Meier estimator was used to estimate the survival function and median time of the patients. Result showed that each patient has a 50% chance of surviving at least 13 months and a minimum of 10 months but not more than 17 months. The Log-rank test was used to test the differences in the survival curves. Result showed significant difference in survival times for International Federation of Gynaecology and Obstetrics (FIGO) stages and recurrence with a p-value of $5e^{-08}$ and 0.007 respectively. Classification and regression tree (CART) was used to predict the chance of survival of the patients. Findings from the CART model revealed that the model had 82.5% of correctly classifying cervical cancer patients and also showed menopause as the most important predictor of cervical cancer. Finally, Accelerated Failure Time (AFT) model was used to determine the prognostic factors associated with cervical cancer. AFT models used were Exponential. Weibull, Log-logistics and Lognormal distributions and based on Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), all models were compared for performance. The Log normal with the minimum AIC and BIC values of 1821.70 and 1844.66 was the best model for the data and was subsequently used for further analysis. Results obtained establish that recurrence and patients who had adenosquamous (ADQ) histological type significantly prolonged the survival time of patients while parity, occupation house wife (H/W), tribe (Yoruba), tumour grade well differentiated (WD) and treatment received (chemotherapy) significantly shortened survival time of patients. The findings of this study showed that Lognormal AFT model described the survival time of the cervical cancer patients dataset better than other distributions used. Furthermore, the study found a high percentage diagnosed at advanced stage, which had negative effect on survival and stressed the need for improving early detection.

TABLE OF CONTENTS

Content **Page**

Cover Page

Title page i

Declaration ii

Certification iii

Dedication iv

Acknowledgements v

Abstract vii

Table of Contents viii

List of Tables xv

List of Figures xvii

List of Appendices xix

CHAPTER ONE

1.0 INTRODUCTION 1

1.1 Background to the Study 1

1.2 Statement of the Research Problem 3

1.3 Aim and Objectives of the Study 4

1.4 Justification for the Study 5

1.5 Scope and Limitation of the Study 5

1.6 Definition of Operational Terms 5

CHAPTER TWO

2.0 LITERATURE REVIEW 7

2.1 Preamble 7

2.2 Cancer Stages 7

2.3 Cancer Treatment 9

2.4 Histologic Type 11

2.5 Metastasis 12

2.6 Survival Analysis 12

2.6.1 Objectives of survival analysis 13

2.6.2 Challenges of survival analysis 13

2.6.3 Survival data 14

2.6.4 Survival time 14

2.7 Censoring 15

2.7.1 Data censoring 15

2.7.2 Assumptions of censoring 16

2.7.2.1 Independent censoring 16

2.7.2.2 Random censoring 16

2.7.2.3 Non-informative censoring 16

2.7.3 Types of censoring 17

2.7.3.1 Right censoring 17

a) Type I right censoring 17

b) Type II right censoring 18

c) Type III right censoring 19

2.7.3.2 Left censoring 19

2.7.3.3 Interval censoring 20

2.7.4 Validity of censoring 21

2.8 Survival Analysis Techniques (Models) 21

2.8.1. Non-parametric models 22

2.8.1.1 Kaplan-Meier method 22

i Types of Kaplan-Meier survival curves 22

a) Overall survival curve 22

b) Disease free survival curves 22

c) Progression free survival curves 23

d) Disease specific survival curves 23

ii Assumptions of Kaplan-Meier 23

2.8.1.2 Log-Rank test 23

2.8.2 Semi-parametric models 24

2.8.2.1 Cox proportional regression 24

2.8.3 Parametric models 24

2.8.3.1 Weibull parametric model 24

2.8.4 Machine learning 25

2.8.4.1 Classification and regression tree (CART) model 25

2.9 Application of Survival Analysis 25

2.9.1 Survival analysis in engineering 26

2.9.2 Survival analysis in social science 26

2.9.3 Survival analysis in medicine 27

2.10 Research Gap 32

CHAPTER THREE

3.0 RESEARCH METHODOLOGY 33

3.1 Data Source 33

3.2 Data Collection 33

3.3 Study Variables 34

3.4 Data Processing and Analysis 34

3.4.1 Descriptive statistics 34

3.4.2 Kaplan-Meier (K-M) analysis 35

3.4.2.1 Mathematical formulation of Kaplan-Meier model 35

3.4.2.2 Assumptions of K-M estimator 36

3.4.2.3 Algorithm of K-M estimator 36

3.4.2.4 Computation of 95% confidence interval (CI) 37

3.4.3 The Log-Rank test 37

3.4.3.1 Assumptions of the Log-Rank test 38

3.4.3.2 Algorithm for the Log-Rank test 38

3.4.4 Classification tree 39

3.4.4.1 Splitting criterion and impurity measurement 40

3.4.4.2 Receiver operating characteristic (ROC) curve 41

3.4.4.3 Variable importance 42

3.4.4.4 Confusion matrix 42

3.4.5 Accelerated failure time model 43

3.4.5.1 Mathematical formulation of AFT 43

3.4.5.2 Weibull AFT model 45

3.4.5.3 Exponential AFT model 46

3.4.5.4 Log-normal AFT model 46

3.4.5.5 Log-logistic AFT model 47

3.4.5.6 Parameter estimation of accelerated failure time model 48

3.4.5.7 Model selection 49

CHAPTER FOUR

4.0 RESULTS AND DISCUSSION 50

4.1 Descriptive Statistics 50

4.2 Kaplan-Meier (K-M) Analysis 54

4.2.1 Overall K-M analysis for cervical cancer patients 54

4.2.2 Estimate of survival base on marital status 55

4.2.3 Estimate of survival base on religion 56

4.2.4 Estimate of survival base on tumour stages 57

4.2.5 Estimate of survival base on occupation 59

4.2.6 Estimate of survival base on recurrence 60

4.2.7 Estimate of survival base on family history 61

4.2.8 Estimate of survival base on smoking status 62

4.2.9 Estimate of survival base on alcohol consumption 63

4.2.10 Estimate of survival base on tribe 64

4.2.11 Estimate of survival base on comorbidity 65

4.2.12 Estimate of survival base on histology 66

4.2.13 Estimate of survival base on tumour grade 67

4.2.14 Estimate of survival base on treatment 68

4.3 Log-Rank Analysis 70

4.4 Classification and Regression Tree (CART) Model 73

4.5 Accelerated Failure Time Model Fitting 79

CHAPTER FIVE

5.0 CONLUSIONS AND RECOMMENDATIONS 83

5.1 Conclusions 83

5.2 Recommendations 83

5.3 Contribution to knowledge 84

REFRENCES

LIST OF TABLES

Table	Title	Page
2.1	FIGO staging for cervical cancer	7
2.2	Treatment algorithm for cervical cancer	10
3.1	2×2 confusion matrix for CART model	42
4.1	Baseline characteristics of 388 cervical cancer patients for quantitative variables	51
4.2	Baseline characteristics of 388 cervical cancer patients for qualitative variables	52
4.3	K-M estimate of the overall survival function	55
4.4	K-M estimate of survival time base on marital status	56
4.5	K-M estimate of survival time base on religion	57
4.6	K-M estimate of survival time base on tumour stages	59
4.7	K-M estimate of survival time base on occupation	60
4.8	K-M estimate of survival time base on recurrence	61
4.9	K-M estimate of survival time base on family history	62
4.10	K-M estimate of survival time base on smoking status	63
4.11	K-M estimate of survival time base on alcohol consumption	64
4.12	K-M estimate of survival time base on tribe	65
4.13	K-M estimate of survival time base on comorbidity	66

4.14 K-M estimate of survival time base on histology 67

4.15 K-M estimate of survival time base on tumour grade 68

4.16 K-M estimate of survival time base on treatment 69

4.17 Log-Rank test of difference 70

4.18 CART model confusion matrix 73

4.19 CART model summary 75

4.20 Optimal tree and variable importance summary 79

4.21 Table 4.21: Log-likelihoods, AIC and BIC in the models 80

4.22 Estimate of prognostic factors of cervical cancer based on NHA 80

LIST OF FIGURES

Figures	Title	page
2.1	Right censoring	17
2.2	Type I right censoring	18
2.3	Type II right censoring	18
2.4	Type III right censoring	19
2.5	Left censoring	20
2.6	Interval censoring	20
2.7	Taxonomy of survival analysis methods	21
4.1	Overall survival curve for cervical cancer patients	55
4.2	Estimate of survival curve base on marital status	56
4.3	Estimate of survival curve base on religion	57
4.4	Estimate of survival curve base on tumour stages	58
4.5	Estimate of survival curve base on occupation	60
4.6	Estimate of survival curve base on recurrence	61
4.7	Estimate of survival curve base on family history	62
4.8	Estimate of survival curve base on smoking status	63
4.9	Estimate of survival curve base on alcohol consumption	64
4.10	Estimate of survival curve base on tribe	65
4.11	Estimate of survival curve base on comorbidity	66

4.12 Estimate of survival curve base on histology 67

4.13 Estimate of survival curve base on tumour grade 68

4.14 Estimate of survival curve base on treatment 69

4.15 Receiver operating characteristic (ROC) curve 74

4.16 Optimal tree diagram 76

4.17 Relative variable importance 78

LIST OF APPENDICES

Appendix	Title	Page
I	Misclassification	95
II	Model selection based on AICc	96
III	Output of insignificant variables from Log normal analysis	97
IV	Testing the Cox proportional assumption	99
V	R Codes	102

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background to the Study

Cancer is a generic term used for a group of diseases that cause abnormal cells to divide without control and over pass other tissues (Ahmed *et al*., 2015 and Tefsay *et al*., 2021). Expansion out of control of these cells can result in death (American Cancer Society, ACS, 2014). It is the leading cause of death worldwide accounting for 7.6 million deaths (about 13% of all deaths) in 2008 and is projected to continue rising with an estimate of 13.1 million deaths in 2030 (Ferlay *et al*., 2010). According to (Bray, 2014), an estimate of over 20 million new cancer cases is expected annually by 2025.

One of the areas plagued with the prevalence of non-communicable diseases, especially cancers which is on the increase is sub-Saharan Africa. World Health Organization (WHO, 2018) report attests to the frightening dearth of a working cancer control plans in sub-Saharan countries despite the enormity of socioeconomic disturbance it present to the countries. Only 17% of African countries have sufficiently financed cancer control program and according to (WHO, 2019) none is within sub-Sahara region. Akinde *et al*. (2015) reported that in Nigeria, about ten thousand (10,000) cancer casualties and two hundred and fifty thousand (250,000) new cases are registered every year. The high case fatality rate of cancer in Nigeria is due to low level of cancer awareness and screening, late discovery, unhealthy lifestyle, superstitious beliefs, limited or poorly funded healthcare facilities and dearth of oncology experts (Adamu *et al*., 2019, Ilevbare *et al*., 2020).

Cervical cancer is recognised as the most common gynaecological malignancy with more than half a million women diagnosed in 2012 (Torre *et al*., 2015). In sub-Saharan Africa, (Akinremi *et al*., 2005 and Awodele *et al*., 2011) reported cervical cancer to be a major public health menace to women of all age groups and cancer of any type is viewed as an unwelcome guest in every home, and it is often seen as a death sentence in Nigeria (Adamu *et al*., 2019). This cancer is caused by human papilloma virus (HPV) infections (Goldie *et al*., 2003 and De Sanjose *et al*., 2010) with nearly one third of women succumbing to the disease in the first five years following diagnosis (National Cancer Institute, 2018). Kidanto *et al*. (2002), Mwaka *et al*. (2016), Elmajjaoui *et al*. (2016) and Gizaw *et al*. (2017) reported that women with cervical cancer in sub-Saharan Africa are frequently diagnosed at advanced stages. They also have poor access to appropriate diagnosis and treatment (Denny, 2011) and prolonged treatment waiting times (Kantelhardt *et al*., 2014). All these factors contribute to poor survival outcomes (Denny, 2011; Kantelhardt *et al*., 2014; Elmajjaoui, 2016; Gizaw, 2017).

Jedy-Agba, *et al*. (2012) reported cervical cancer as the most common cancer in women in Nigeria. Nigeria is rated tenth in cervical cancer mortality worldwide (Okoye, 2014). In a study carried out by (Anorlu *et al*., 2010) at the gynaecology ward of Lagos University Teaching Hospital from 2002 to 2007, 104 casualties were registered and 83 (79.8%) of these were traceable to cancer. Furthermore, Anya *et al*. (2006) reported cervical cancer to be the leading cause of deaths from 2033 patients at the gynaecology ward in Enugu followed by choriocarcinoma, septic abortion and ovarian cancer in that order.

Survival is the complex outcome of number of agents such as the availability of screening programmes, treatment infrastructure, stage at diagnosis, socioeconomic factors and healthcare workers to give well timed and appropriate care (Kidanto *et al*., 2002; Denny, 2011; Kantelhardt *et al*., 2014 and Elmajjaoui, 2016). Data from records of registered cancer cases provide researchers with an opportunity to calculate survival estimates which approximate real life scenarios in a given study area. These records helps in monitoring population-based indicators of cancer control such as incidence, survival and mortality (Piñeros *et al*., 2017) which in turn guides in evidence-based cancer policy formulation (Parkin, 2006). Often in cancer studies, the main outcome under assessment is the time to an event of interest (Clark *et al*., 2003).

Survival analysis studies employs the use of the technique of "censoring" whereby each patient contributes data even if he/she does not achieve the desired outcome of interest or drops out during the cause of the study for any reason (Gogtay and Thatte, 2017). Moreover, the defining characteristic of survival studies is that the response variable could be censored; in other word, there may be no definitive event time for some patients (Yoosefi *et al*., 2018). This is done because time to an event is a far more crucial tool than looking at the event alone as studying how patients respond to treatment over time are very important to understanding how treatment influences both disease progression and quality of life (Gogtay and Thatte, 2017). Most oncology research with follow-up studies are characterised by evident outcomes such as death, recurrence and response to treatment. Time-to-event outcomes are distinctly noticed in the follow-up study using either cohort or clinical trial study designs (Perera and Dwivedi, 2019). Besides incidence and clinical stage at presentation, knowledge of survival is essential to monitor and evaluate cancer control programme.

1.2 Statement of the Research Problem

Cervical cancer is a major cause of morbidity and mortality among women and continues to have significant influence on women worldwide, particularly women in developing countries (Vishma *et al*., 2016). This cancer placed among the top four female cancers with world-wide forecast incidence of 569,847 novel cases in 2018 corresponding to 15.1

new cases per 100,000 women and a cumulative risk of 1.36% from birth to 75 years old (Anfinan and Sait, 2020). This cancer constitute one of the leading cause of cancer affiliated death rate in females, accounting for 311,365 global mortalities in 2018. In Nigeria, 70,327 deaths in women have been attributed to cancer with cervical cancer causing 14.89% of those deaths in 2018 making it second most common cancer after breast cancer (Anyasi and Foss, 2021). A frequent occurrence and lower survival rate in Nigeria is a trend of late presentation and diagnosis at advanced stages of the disease, lack of awareness, inadequate screening facilities, inadequate treatment due to poor patient compliance, financial constraints and inadequate treatment facilities leading to poor prognosis. Consequently, this research work intends to investigate the survival rates and the prognostic factors of cervical cancer patients.

1.3 Aim and Objectives of the Study

The aim of the study is to fit a parametric survival model that best describe the survival time of patients with cervical cancer with respect to the predictors considered. The specific objectives are: To

i. Perform a descriptive analysis to determine the demographic, social and clinical characteristics of the patients.
ii. Estimate the survival function and median time of patients with cervical cancer.
iii. Investigate the difference in survival time of the patients in different groups.
iv. Predict the chance of survival of a cervical cancer patient using the classification and regression tree (CART) model.
v. Fit the parametric model and identify the prognostic factors associated with cervical cancer.

1.4 Justification for the Study

Cervical cancer is a disease of great public interest as it primarily affect women aged 15-44years, an age group within which women make great social and economic contribution to the society. The main cause of cervical cancer is persistent infection with a sexually transmitted virus “human papilloma virus” (HPV). Various analyses of application of survival analysis have been carried out on cervical cancer, but only a few have been carried out to assess the prognostic factors of the disease, and none was in Nigeria. Furthermore, the few studies use cox proportional hazard which is a semi parametric model. Therefore, this research work intends to a parametric survival model to identify the prognostic factors for cervical cancer.

1.5 Scope and Limitation of the Study

The scope of this study will be limited to the use of secondary data collected from National Hospital, Abuja, Nigeria, for the period 2010 to 2020. It is further subjected to only available dataset as at the period stated for time and financial situations.

1.6 Definition of Operational Terms

Survival time: The time origin to the start of outcome of interest.

Time to failure: The interval of time from beginning of an event to the outcome of interest.

Censoring: a scenario whereby the entire survival times are not known.

Hazard: The rate at which a particular event occurs.

Hazard ratio: The ratio of the hazard in the experiment to the hazard in the control arm.

Survival function S(t): The probability that an individual survives from a specified time point to a specified future time, where $0 \leq S(t) \leq 1$.

Hazard function h(t): The probability that an individual who is under an observation at a time t, has an event at that time.

Covariates: variable that is possibly predictive of the outcome under study.

Parity: The number of pregnancies that attain delivery age, mostly between 5-7 months.

Menopause: The final menstrual period of a woman after which ovulation no longer occurs.

Menarche: The onset of menstruation, a girl's first period.

Coitarche: The first sexual intercourse.

Nullipara: A woman who has never given birth.

Histopathology: The microscopic study of tissue, especially of abnormal tissue as a result of disease.

Histology: The study of the microscopic structure, chemical composition and function of the tissue or tissue systems of animals.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Preamble

Cervical cancers are malignant procreation that develop from the cervix (Anfinan and Sait, 2020). Majority of all cervical cancer cases (99%) are associated by infection with high-risk sexually transmitted human papillomaviruses (HPV). In addition to the infection of HPV, (Schiffman *et al*., 1996; Bosch *et al*., 1999; Franco *et al*., 1999; Walboomers *et al*., 1999; IARC, 2007) reported early age at coitarche, numerous childbirth, several sexual partners and smoking as other risk factors.

2.2 Cancer Stages

The international federation of gynaecology and obstetrics (FIGO) is used in cervical cancer staging (Abd Razak, 2016). According to (Waggoner, 2003), cancer stage is determined clinically and what is considered is mainly the size of the tumour or its extension into the pelvis. Description of cancer stage is presented in Table 2.1.

Table 2.1: FIGO staging for cervical cancer (Waggoner, 2003)

Stage	Description
Stage 0	Carcinoma in situ, cervical intraepithelial neoplasis grade III
Stage I	The carcinoma is strictly confined in the cervix (extension to the corpus would be disregarded)
Ia	Invasive carcinoma which can be diagnosed only by microscopy. All microscopically visible lesions- even with superficial invasion – allotted to stage Ib carcinoma. Invasion is limited to a measured

Table 2.1: Continued

	stromal invasion with a maximal depth of 5.0mm and a horizontal extension of not >7.0mm depth of invasion should not be >5.0mm taken from the base of epithelium of the original tissue should not change the stage allotment.
Ia1	Measured stroma invasion of not >3.0mm in depth and extension of not >7.0mm
Ia2	Measured stroma invasion of not >3.0mm and not >5.0mm with an extension of not >7.0mm
Ib	Clinically viable lesions limited to the cervix uteri or preclinical cancers greater than stage Ia
Ib1	Clinically visible lesions not >4.0cm
Ib2	Clinically visible lesions >4.0cm
Stage II	Cervical carcinoma invades beyond uterus but not to the pelvic wall or the lower third of the vagina.
IIa	No obvious parametrial involvement
IIb	Obvious parametrial involvement
Stage III	The carcinoma has extended to the pelvic wall. On rectal examination, there is no cancer-free space between the tumour and the pelvic wall. The tumour involves the lower third of the vagina. All cases with hydronephrosis or non-functioning kidney are included unless they are known to be due to other causes.

Table 2.1: Continued

IIIa	Tumour involves lower third of the vagina, with no extension to the pelvic wall.
IIIb	Extension to the pelvic wall/or hydronephrosis or non-functioning kidney
Stage IV	The carcinoma has extended beyond the true pelvic or has involved (biopsy proven) the mucosa of the bladder or rectum. A bullous edema, as such, does not permit a case to be allotted to stage IV.
IVa	Spread to the growth to adjacent organs.
IVb	Spread to distant organs.

2.3 Cancer Treatment

Cancer treatment options are influenced by several factors such as age, general condition of the patients, the stage of the tumour and patients own preference (Radstone and Kunkler, 2003). Patients with early stage cancer will be treated by radical hysterectomy and pelvic lymphadenectomy or alternatively combined external pelvic irradiation and brachytherapy with concomitant chemotherapy (Jensen, 2007). Meanwhile patients with late or more advanced cancer will be given a combination of radiotherapy and concomitant chemotherapy (Abd Razak, 2016). Table 2.2 shows types of treatment for cervical cancer patients.

Table 2.2: Treatment algorithm for cervical cancer (Waggoner, 2003)

Stage	Clinical features	Treatment
IA1	Invasion 3.0mm or less	If patient desires fertility, conisation of cervix If she does not, simple hysterectomy (abdominal or vaginal).
IA2	With lymphatic space invasion	Hysterectomy with or without pelvic lymphadenectomy
IB1	3.0-5.0mm invasion, <7.0 mm lateral spread	Radical hysterectomy with pelvic lymphadenectomy Radiotherapy.
IB2	Tumour 4cm or less	Radical hysterectomy with pelvic lymphadenectomy plus chemoradiotherapy for poor prognostic surgical-pathological factors*. Radiotherapy
IIA	Tumour bigger than 4cm	Radical hysterectomy with pelvic lymphadenectomy plus chemoradiotherapy for poor prognostic surgical and pathological factors*. Chemoradiotherapy. Chemoradiotherapy plus adjuvant hysterectomy.

Table 2.2: Continued

IIB	Upper-two-thirds vaginal involvement	Radical hysterectomy with pelvic lympadenectomy. Chemoradiotherapy.
IIIA	With parametrial extension	Chemoradiotherapy.
	Lower third vaginal involvement	Chemoradiotherapy.
IVA	Local extension within pelvis	Chemoradiotherapy. Primary pelvic extenteration.
IVB	Distant metastases	Palliative chemotherapy. Chemoradiotherapy.

*Pelvic lymph-node metastases; large tumour; deep cervical stromal invasion; lymphovascular space invasion; positive vaginal or parametrial margins.

2.4 Histologic Type

The histological type of cervical carcinoma as classified by world health organisation (WHO) are squamous cell carcinoma, adenocarcinoma and other epithelial tumours (Cheah and Looi, 1999) and most prevalent of the three types is Squamous cell carcinoma (SCC). The infection of the high-risk type of HPV is greatly related with the two histologic type (Prendiville and Sankaranayanan, 2017; Walboomer *et al.*, 1999). SCC arise from the squamous cells lining the cervix outer area. Adenocarcinoma begins in the column-shaped glandular cells that lined the cervical canal. Adenosquamous carcinoma, adenoid cystic carcinoma and neuroendocrine carcinoma are the other epithelial tumours that may be present (Abd Razak, 2016).

2.5 Metastasis

Metastasis is the dispersion of cancer cells to remote regions of the body, often by way of the lymph system. The spread could be through direct local extension or through the lymph system (Abd Razak, 2016). Moore Higgs and Chafe (2001) reported that the spread could travel all the way to the vagina mucosa, parametrial tissues and ligaments, endometrial cavity, pelvic side wall, bladder and rectum. According to (Ho *et al*., 2004), para-aortic lymph node and pelvic metastases are the notable prognostic factors of cervical cancer but experience of distant metastasis (spread to remote organs like bladder, lungs and bones) have also been documented.

2.6 Survival Analysis

Measuring the occurrence of an outcome or event are majorly of interest in medical and epidemiological studies (Prinja *et al*., 2010). However, survival analysis studies are centred on evaluating time to an event of interest (Prinja *et al*., 2010 and Atlam *et al*., 2021). Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is time until occurrence of an event (Clark *et al*., 2003; Bradburn *et al*., 2003; Gogtay and Thatte, 2017). According to Chakraborty (2018) survival analysis refers to statistical techniques which have been designed to circumvent the issues arising out of incomplete information regarding the time until which a desired event or end point occurs. It is one of the most significant advancement of mathematical statistics in the last quarter of 20th century and has become the de facto standard in biomedical data analysis (Ma and Krings, 2008). This method of analysis was initially developed to analyse risk of death over time, but is now used for the analysis of many categorical outcomes in health research (Tolley *et al*., 2016). In the real world, survival time could be time to a light bulb fusing, time to replacing the battery on the wall clock or time to change the gas cylinder, however, in medical parlance, this time may range from time to fatal event i.e. death, time to metastases, onset of disease, time to tumour recurrence, time to discharge from the hospital, time to first exacerbation after a new drug treatment in patients with disease, time to dialysis in patients with renal dysfunction, attainment of a biochemical marker and bankruptcy (Prinja *et al*., 2010 and Seungyeoun and Heeju, 2019).

2.6.1 Objectives of survival analysis

The objectives of survival analysis according to (Kartsonaki, 2016) includes:

(i) The analysis of patterns of event times.

(ii) The comparison of distributions of survival times in different groups of individuals.

(iii) Examining whether and by how much some factors affected the risk of an event of interest.

2.6.2 Challenges of survival analysis

Time-to-event data encounters several research challenges such as censoring, symptoms (features) correlations, high-dimensionality, temporal dependencies, and difficulty in acquiring sufficient event data in a reasonable period of time (Atlam *et al*., 2021). Two practical issues are encountered while dealing with survival time. It is difficult to specify the start time in some cases. For instance, assuming the starting time to be the onset of a disease, it would be impossible to correctly establish this time. Secondly, it is difficult to establish failure time. For example, if failure time is recorded as the time of death, that would not cause any difficulty, but in a situation when a subject simply decides that she/he wants to leave the study, or it survives more than the established record time. This type of survival time is called a censored survival time.

2.6.3 Survival data

The study of survival data has focused on predicting the probability of response, survival, or mean lifetime, comparing the survival distributions of experimental animals or of human patients and the identification of risk and/or prognostic factors related to response, survival and the development of a disease (Elisa and John, 2003). Survival data consist of two sets of information for each subject viz the time under observation and the ultimate outcome at the end of that time and features a varied length of follow-up time among participants and also the event of interest is almost never observed by the end of the study (Johnson, 2018). According to (Elisa and John, 2003), many scholars view survival data analysis to be merely the application of two conventional statistical methods to a special type of problem viz-a-viz parametric if the distribution of survival times is known to be normal and nonparametric if the distribution is unknown. Clark *et al*. (2003) reported that this data are rarely normally distributed, but are skewed and comprise typically many early events and relatively few late ones. It is this features of the data that make the special methods called survival analysis necessary (Clark *et al*., 2003).

2.6.4 Survival time

According to (Elisa and John, 2003), survival time is the time to the occurrence of a given event of interest such as death, relapse of disease, response to treatment, unemployment and completion of a task. This interval is usually in days, months or years between the start of follow up for that subject until the occurrence of the event of interest or until censored (Flynn, 2012). This time can be tumor-free time (Anfinan and Sait, 2020 and Chen *et al*., 2021), the time from the start of treatment to response (Sannachi *et al*., 2018), length of remission (Sabbatini and Spriggs, 2006 and Harrison *et al*., 2007), and time to death (Vishma *et al*., 2016; Gurmu *et al*., 2018; Wassie *et al*., 2019). Johnson (2018) reported that in epidemiological study, the time origin may be birth, time of first exposure or another point in time. Analysis of survival time is complicated because the follow-up length is often different for each participant, and the event of interest is not observed in all the subjects by the end of the study (Fadnavis, 2019 and Atlam *et al*., 2021). For participant in whom the event of interest is not observed, what is known is that their survival times are longer than their time spent in the study, but their exact survival times are unknown (Johnson, 2018).

2.7 Censoring

Survival analysis is based on the time measured from a relevant time origin to a particular event of interest. However, the event of interest may not be observed for some patients because of end-of-study censoring, loss to follow-up or competing events (such as deaths from other causes). In these cases the patient's survival is said to be censored since his/her actual survival time is known to be larger than the observed one (Tefsay *et al*., 2021). According to (Flynn, 2012), censoring is an important concept that relates to subjects who form part of a cohort but who never experience the event of interest. In other words it refers to the situations in which exact lifetimes are fully observed for only a portion of the individuals in a statistical population sample (Ma and Krings, 2008).

2.7.1 Data censoring

In survival analysis studies, the survival times of all the subjects are not known leading to a skewed survival distribution or a distribution far from being normal (Elisa & John, 2003). Applying conventional statistical methods often leads to deficient outcome because some subjects in the study may have not experienced the event of interest at the end of the study or time of analysis leading to unknown exact survival times of these subjects (Seungyeoun and Heeju, 2019). These set of subjects that have not experienced the event of interest at the end of the study are referred to as censored observations or censored times and can also occur when people are lost to follow-up after a period of study. According to (Thuijs *et al.*, 2018) censoring mean that the time to the event of interest cannot be determined. Censoring actually makes survival analysis different in the sense that specific difficulties relating to survival analysis arise largely from the fact that only some individuals have experienced the event and subsequently survival times will be unknown for a subset of the study group (Seungyeoun and Heeju, 2019).

2.7.2 Assumptions of censoring

2.7.2.1 Independent censoring

Kleinbaum and Klein (2012) reported independency of censoring to be the most useful for drawing correct inferences that compare the survival experience of two or more groups. This assumption essentially specify that within any subgroup of interest, the subjects are censored at time “t” with respect to their survival experience.

2.7.2.2 Random censoring

This imply that subjects who are censored at time t should be representative of all the study subjects who remained at risk at time “t” with respect to their survival experience.

2.7.2.3 Non-informative censoring

This occur if the distribution of survival time (T) provide no information about the distribution of censorship time (C).

2.7.3 Types of censoring

2.7.3.1 Right censoring

Right censoring is said to occur when despite continuous monitoring of outcome event, the subject is lost to follow up or the event does not occur within the study duration (Stevenson, 2007 and Prinja *et al*., 2010). According to (Gogtay and Thatte, 2017), survival data is right censored where censoring occurs after the patient has entered the study because the participant has left the study for any reason. Emmert-Streib and Dehmer (2019) reported that right censoring occur as a result of

- A subject dropping out of the study
- The study has a fixed time line and the event occurs after the cut-off time
- A patient experiences a different event that makes further follow-up impossible

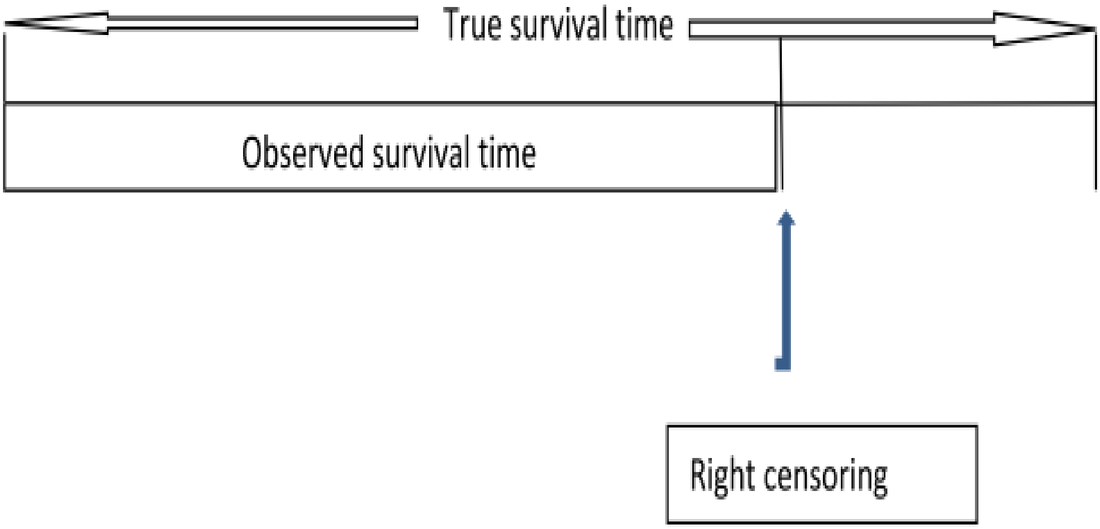


Figure 2.1: Right censoring (Gogtay and Thatte, 2017)

a) Type I right censoring

In this censoring type, all subjects begin and end the study at the same time (fixed length of study) (Emmert-Streib and Dehmer, 2019).

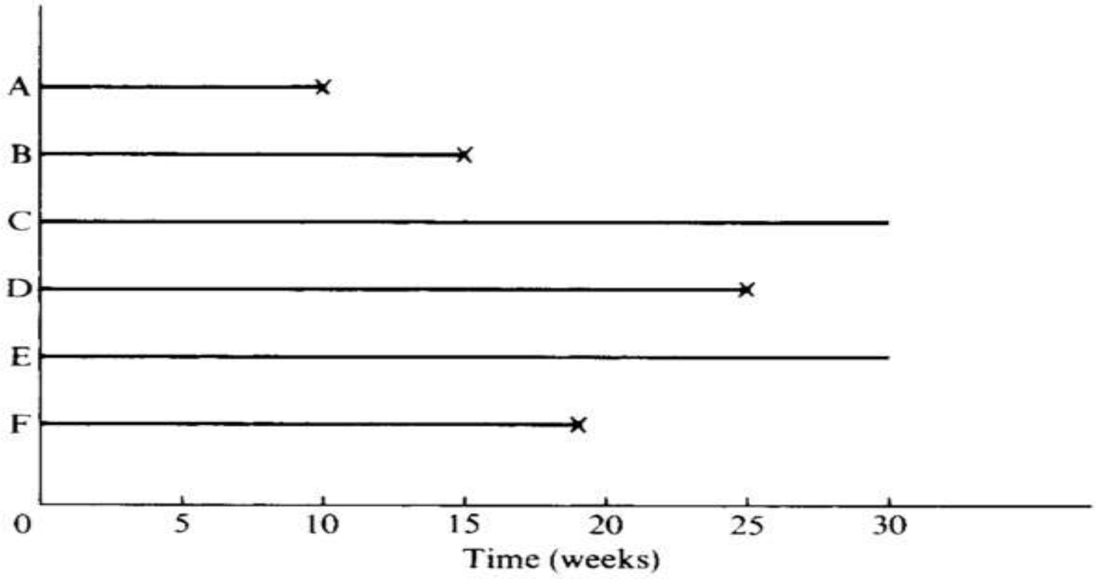


Figure 2.2: Type I right censoring (Elisa and John, 2003)

b) Type II right censoring

In type two right censoring, all subjects begin the study at the same time but the study ends when a predetermined fixed number of subject have experienced the event (flexible length of study).

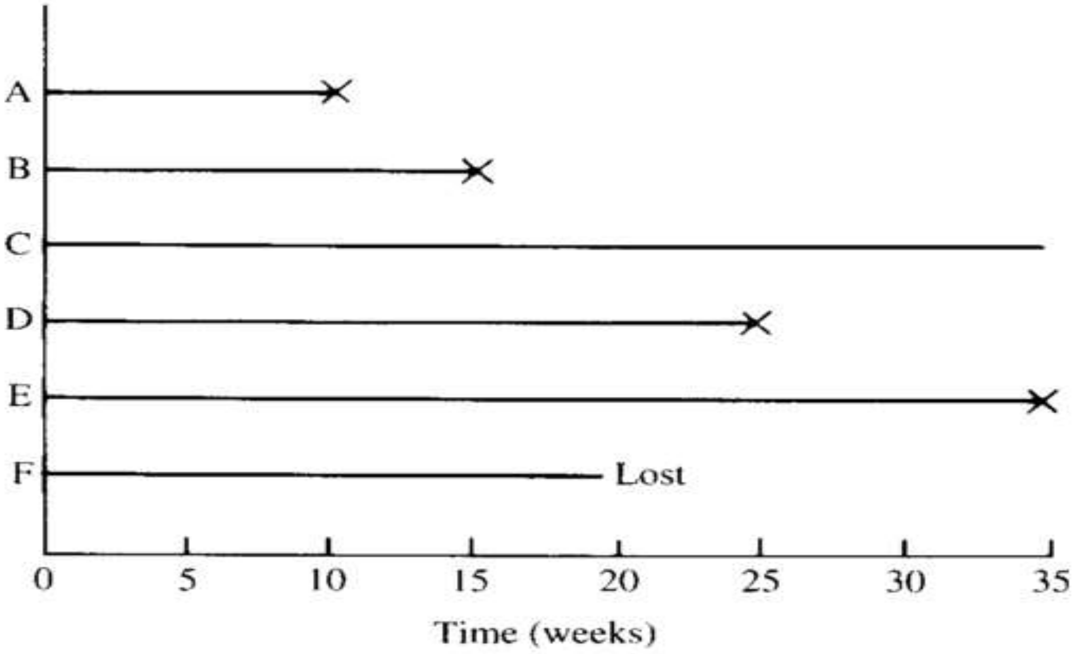


Figure 2.3: Type II right censoring (Elisa and John, 2003)

c) Type III right censoring

Right censoring of type III also called random censoring occurs naturally (Ma and Krings, 2008). In this censoring scenario, the subjects enter the study at different times but the length of study is fixed.

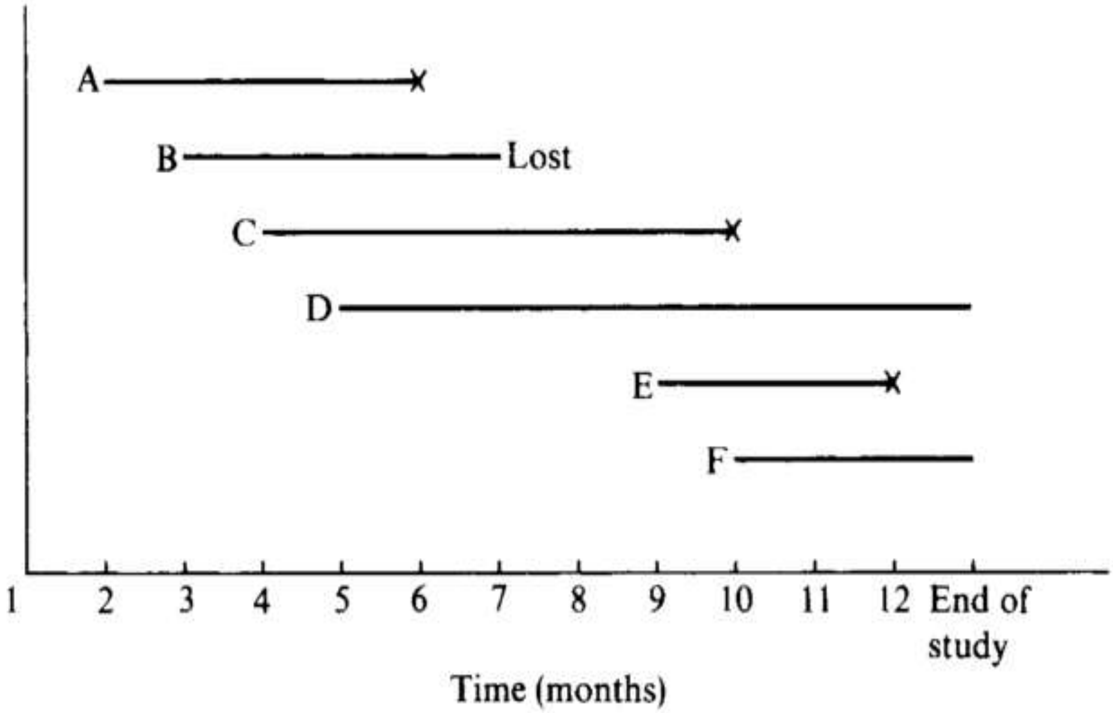


Figure 2.4: Type III right censoring (Elisa and John, 2003)

2.7.3.2 Left censoring

When an event of interest occurred prior to a certain time but the exact time of occurrence is unknown such a situation is termed left censoring (Elisa and John, 2003). According to (Prinja *et al*., 2010), a subject is said to be left censored if the subject had been on risk of ailment for a period before entering the study. In clinical trials, left censoring is usually not a challenge, as starting point is defined by an event such as entry of patient in trial, randomization or occurrence of a procedure or treatment (Chen *et al*., 2021).

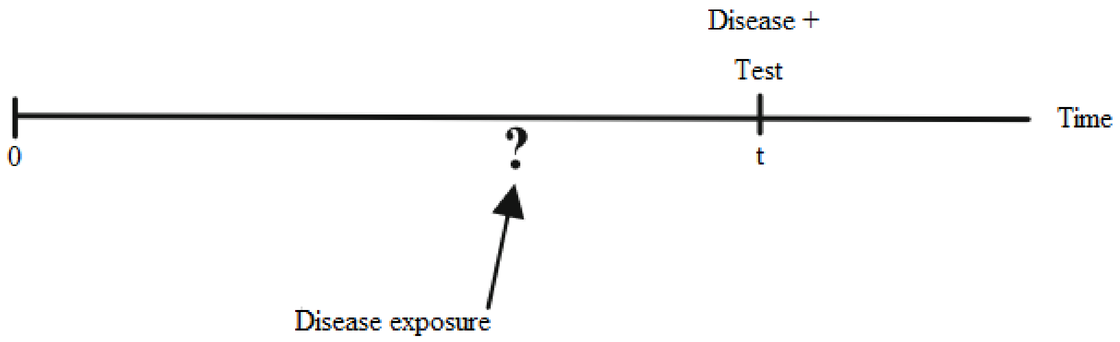


Figure 2.5: Left censoring (Kleinbaum and Klein 2012)

2.7.3.3 Interval censoring

This occur when the event of interest is known to have occurred between two known time intervals (Elisa and John, 2003). Scenario of interval censoring comes up when time to event is known only up to a time interval. This situation occurs in case the assessment of monitoring is done at a periodical frequency (Prinja *et al.*, 2010). Practically, most observational studies dealing with non-lethal outcomes have periodical examination schedules and are thus interval censored. In this censoring type, subjects that are negative on the first test and positive at the next are said to be interval censored with the first sampling date being the lower interval and the second sampling date the upper interval (Stevenson, 2007).

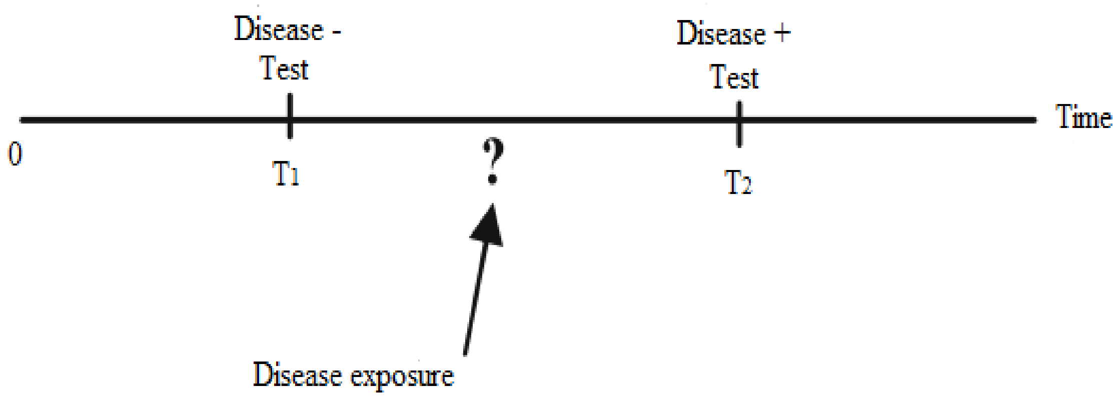


Figure 2.6: Interval censoring (Kleinbaum and Klein 2012)

2.7.4 Validity of censoring

According to (Gogtay and Thatte, 2017), when a subject is censored, the risk of achieving the outcome for the reminder of the subjects who continue on the study should be unchanged. In this case the censoring becomes valid. Also, the censoring should be randomly distributed overtime as its main assumption is that it is independent of time, the intervention/treatment under evaluation.

2.8 Survival Analysis Techniques (Models)

Several models are available for analysing the relationship between a set of predictor variables with the survival time (Gogtay and Thatte, 2017). Statistical methods of survival analysis are mainly classified into parametric, non-parametric and semi parametric methods.

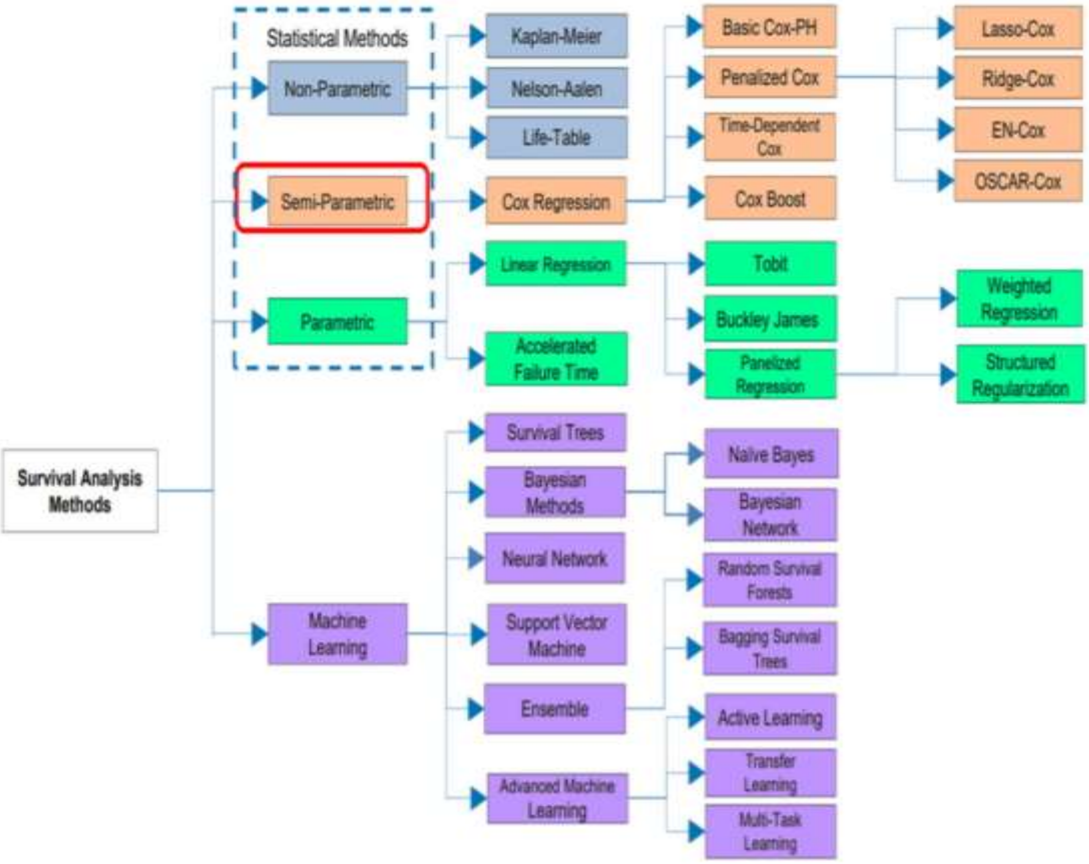


Figure 2.7: Taxonomy of survival analysis methods (Culled from Atlam *et al*., 2021)

2.8.1 Non-parametric models

This models are designed to deal with unobserved heterogeneity and are also good methods to understand basics and produce descriptive results (Mills, 2011). They are the simplest methods which made no assumption about the underlying distribution (or shape) of hazard function (Flynn, 2012).

2.8.1.1 Kaplan-Meier method

The Kaplan-Meier is a widely used non-parametric method for analysing survival data (Gogtay and Thatte, 2017). This method is based on individual survival times and assumes that censoring is independent of survival time, that is, the reason an observation is censored is unrelated to the cause of failure (Stevenson, 2007). The Kaplan-Meier technique employs the use of curves in its analysis. Gogtay and Thatte (2017) reported that these curves are generated at the end of the survival analysis and after calculation of survival probabilities.

i. Type of Kaplan-Meier survival curves

a) Overall survival curve

The outcome of interest in overall survival curves is usually death from any cause, that is, all cause death. This provides a very broad general sense of the mortality of the group (Rich *et al*., 2010).

b) Disease free survival curves

Here, the event of interest is relapsed of a disease rather than death. This is because patients may have relapsed but not yet died. These curves are lower than overall survival curves.

c) Progression free survival curves

According to Rich *et al*. (2010), these curves are employed in isolating and gauging the effect of a particular treatment on an ailment. These survival curves indicate disease progression as a terminal-point that is tumour growth or spread.

d) Disease specific survival curves

The terminal point utilised by these curves is usually death from disease of interest. In these curves, outcomes are limited only to death from a specific disease and thus can be misleading in that it will always be higher than overall survival and disease free survival curves. Here, subjects that have relapse, die from disease related factor (treatments) or die from non-related causes are excluded as events (Rich *et al*., 2010).

ii. Assumptions of Kaplan-Meier

- At any time, subjects who are censored have the same survival prospects as those who continue to follow on the study (non-informative).
- Survival probabilities are the same for subjects regardless of the time point at which they enter the study.
- The event of interest happens at the time or time interval specified.

2.8.1.2 Log-Rank test

The Log-Rank (LR) test is a large – sample chi-square (χ^2) test that uses as its test criterion a statistic that provides an overall comparison of the K-M curves being compared (Kleinbaum and Klein, 2012). It is used to test whether the difference in survival times between two or more groups is statistically different when testing the null hypothesis of the significance. The null hypothesis (H_o) states that the population does not differ in the probability of an event at any time point. Therefore, in this case, the LR test is used to compare these groups if they are equal at any time in the probability of an event.

2.8.2 Semi-parametric models

The semi-parametric models make assumptions about the impact of variables on the outcome but not the study of the hazard function (Flynn, 2012).

2.8.2.1 Cox proportional regression

The Cox proportional hazard regression of the semi-parametric methods is the commonly used semi-parametric models (Prinja *et al*., 2010). According to Gogtay and Thatte (2017), the Cox proportional hazards models is capable of addressing heterogeneity due to unobserved effects. The model allows the data to determine the baseline hazard (Mills, 2011).

2.8.3 Parametric models

The parametric methods assume that the survival time conforms to some specific statistical distributions (Gogtay and Thatte, 2017). It make assumptions about the impact of variables on outcomes and the shape of the hazard function (Flynn, 2012). Considering their flexibility and variety in function and performance, parametric models are of particular interest to many researchers (Yoosefi *et al*., 2018).

2.8.3.1 Weibull parametric model

This model is known for its advantage of considering that the population may not be homogeneous and estimate parameters by finding values that maximise the likelihood function (Haughton and Haughton, 2011).

2.8.4 Machine learning

2.8.4.1 Classification and regression tree (CART) model

The classification and regression tree (CART) analysis is a parametric and non-linear approach based on recursive partitioning analysis (Marshall, 2001). It is an innovative decision tree algorithm in which several ‘predictor’ variables are crucial to identify patients at different levels of risk through the progressive binary splits (Takahashi *et al.*, 2006). CART can handle numerical data that are highly skewed or multi-modal as well as category predictors with either ordinal or non-ordinal structure (Tittonell *et al.*, 2008).

2.9 Application of Survival Analysis

Survival analysis has found application in many fields such as medicine, biology, engineering, business, arthropology, criminology, economics, epidemiology, social and behavioral sciences (Elisa and John, 2003; Emmert-Streib and Dehmer, 2019; Seungyeoun and Heeju, 2019; Ma, 2021). Depending on the field where it is used, survival analysis is also known as lifetime data analysis, time to event analysis, reliability or event history analysis (Prinja *et al.*, 2010). This analytical method is suitable for description of survival of a single group of subjects, but more interestingly used to compare the experiences of different groups of subjects (Flynn, 2012). The analysis is widely used in clinical and epidemiological research. In randomised clinical trials, it is used to compare the occurrence of outcomes in patients receiving different treatments to establish the most effective treatment (Dumville *et al.*, 2009 and Severe *et al.*, 2010). According to Versmissen *et al*. (2008) and De Oliveira *et al*. (2010), observational (non-randomised) research also makes extensive use of survival models to determine and test the existence of epidemiological association. The output of survival analysis can take the form of life tables, survival curve, formal hypothesis tests and measures of relative risk (Flynn, 2012).

2.9.1 Survival analysis in engineering

Ozturk *et al*. (2018) investigated the potential factors affecting wind turbine failure and model the hazard rate of wind turbine using survival analysis considering operational, climatic and geographic factors. It was discovered that adequately scheduled maintenance can increase the survival of wind turbine systems and electrical subsystems up 2.8 and 3.8 times respectively compared to the systems without schedule maintenance. In another study by (Mohammedi *et al*., 2020) which explore the changes in failure rates of network elements after implementing an intermittent water supply (IWS), it was discovered that the probability of failure rates significantly increase after implementing the IWS scheme, and hence remain for several years after, even when the network returns to continuous water supply (CWS). Ghodrati and Uday (2005) examine the effect of operating environment on the reliability characteristics of component. The result indicated the operating environment of system/machine has considerable influence on system performance. Anto and George (2019) assess lifetime of aircraft glass using lognormal survival model. Findings showed the total amount of risk to aircraft glass failure until 48.83 months is 65%.

2.9.2 Survival analysis in social science

Using survival analysis models, Ayaneh *et al*. (2020) conducted a study to estimate the time spell to first employment and to ascertain the effects of related factors on the time taken to first employment on new graduates. Result from the study shows a median time to first employment of graduates to be 15 months and they concluded that 50% of the graduates were able to secure their first job by 15 months after their graduation date.

Daraba *et al*., (2017) examine the impact of individual characteristics on the length of unemployment spell using Kaplan-Meier and Cox regression models. Findings from their work shows that unemployment benefit, unemployed category and residential environment had significant influence on unemployment spells.

Potential determinants affecting dissolution of marriage were studied by Sanizah *et al*. (2014) using survival analysis approach. Result from their work shows that age at marriage of husband and attending counselling sessions significantly affect marital dissolution. In another study by (Frempong *et al*., 2012), the independence of type of marriage (customary and ordinance) contracted and divorce was tested. Findings reveals that divorce was time dependent but independent of the type of marriage contracted and time for survival was found to be 5 years after marriage.

Mavri and Ioannou (2008) examined customer switching behaviour in Greek banking services to investigate predictors of churn behaviours as part of customer relationship management (CRM). The findings reveal that quality of the offered banking products and services in combination of bank's brand name have a positive effect on the decrease of switching behaviour while demographic characteristics had minimal impact. Kitabo and Kim (2014) studied factors that could contribute in lifting-up the loan repayment rate of customers of commercial bank. The findings reveal that educational level, having previous loan experience, mode of repayment, collateral type and purpose of loan are significantly related with loan repayment rate of customers.

2.9.3 Survival analysis in medicine

Over the years, survival analysis has found great application in the medical parlance. Gurmu (2018) reported the potential risk factors affecting survival time of women with cervical cancer with death as the outcome variable. Findings from the study suggested age, smoking, stage, family history, abortion history, living with HIV/AIDS, age at first marriage and age at first birth had major influence on survival time of the patients. Mascarello *et al*. (2013) and Anfinan and Sait (2020) analysed survival data of women with cervical cancer and also described associated prognostic factors. Result reveal that early stage diagnosis and treatment are key to reducing mortality from cervical cancer. Similarly, Musa *et al*., (2016) reported a retrospective study on 65 invasive cervical cancer cases with a prospective follow up to ascertain the time from diagnosis to mortality among the subjects. All-cause mortality was the event of interest. Findings reveal that advance stage disease and base line anaemia were independently associated with higher death rate. In another retrospective study conducted by (Wassie *et al*., 2019) to evaluate survival status and associated factors of deaths among cervical cancer patients, it was discover that the overall survival rate was lower than in high and middle-income countries, and factors associated with death were advance stage, advance age, comorbidity, base line anaemia and treatment modality. Vishma *et al*. (2016) conducted a combined prospective and retrospective study to determine the survival rate and prognostic factors for 380 cervical cancer patients. The result showed five year survival for cervical cancer to be 48%. The prognostic factors for the disease were age at diagnosis, performance status at presentation, staging and treatment duration. Meanwhile, (Yagi *et al*., 2019) evaluated the trend of cervical cancer in japan using multiple imputation method to estimate age adjusted incidence, relative survival and conditional survival rates. Findings from the study reveals age to be an important predictor of radiotherapy resistance in cervical cancer.

Furthermore, (Carter *et al*., 2021) reviewed the records of 337 confirmed cases of tuberculosis patients in Monrovia and examine the risk factors affecting the survival and multidrug-resistance tuberculosis patients. They concluded that early intervention is required on local tuberculosis, increase in public awareness, and improvement be made on control factors that may affect the survival and multidrug-resistance of tuberculosis patients. In the same vein, (Bolarinwa and Micheal, 2020) reported a survival analysis study on tuberculosis data using time to recovery from TB infection as outcome variable. Result suggested that age, gender and occupation were the major elements of recovery period of TB patients.

Yu *et al*. (2021) explore the survival effect of radiotherapy in ovarian cancer using Kaplan-Meier and Log-Rank tests with overall survival (OS) and cause specific survival (CSS) as end points. Findings shows radiotherapy was associated with a poor prognosis regardless of pathology or cancer stage.

Examination of possible prognostic factors that may affect the survival of breast cancer patients using Weibull parametric model was done by (Ahmad *et al.*, 2015). Findings show that patients with lymphovascular invasion were at 2.13 time greater risks of death due to breast cancer. Impact of socioeconomic, demographic, environmental, health related and nutritional factors in under-five mortality of child was evaluated by (Saroj *et al.*, 2018). Results reveal women age, parity, birth in last five years, number of children alive, birth order and delivery by caesarean section to be statistically significant on child survival.

In yet another study by (Eryurt and Koc, 2012) to compare the fertility behaviour of migrants with those of non-migrants at both origin and destination areas. It was discovered that rural natives and rural-to-rural migrant women experience all the events related with family formation earlier in their life cycle.

The effective methods to prolong the survival of colorectal cancer patients and determine the influential factors in their longevity was studied by (Yoosefi *et al.*, 2018). The result shows the mean survival rate to be 4.52 ± 0.182 years and age at diagnosis as the only significant influential factor. Findings by (Salina-Escudero *et al*., 2020) on a research to ascertain the factors associated with COVID-19 deaths in Mexican population using Kaplan-Meier curve and Cox proportional hazard model show that risk of dying at any time during follow-up was higher among men, older age groups, people with chronic kidney disease and people hospitalised in public health centres. Meanwhile Adamu *et al*. (2019) assess the survivorship time of the real data of cancer patients in the North-Eastern Nigeria and negative impact of insurgency on the life expectancy of its inhabitants. Result revealed high incidence of cancer and reduction in probability of survival as the survival time increases.

In a study conducted by Cao *et al*. (2015) to assess factors affecting the survival time of patients with pancreatic cancer, the outcome of the research showed that Karnofoky performance scale (KPS) was a significant prognostic factor of pancreatic cancer and spleen-invigorating compounds could also have an effect on improving the prognosis of pancreatic cancer patients. Fagbamigbe and Idemuda (2016) model timing of first child birth among Nigerian women and also ascertain socio-demographic and other factors affecting its timing. Findings showed early first birth and age at first marriage to be influenced or dependent on the level of education of women. Also delay of first child birth will be achievable with quality education at early stages in life. In a study conducted by (Khaemba *et al*., 2013) to estimate the cure fraction, the survival time, survival rate and identify covariates that significantly affect the survival of patients with cervical cancer, poor survival rates among patients with distant metastasis and hence increased risked of death compared to those with localised cancer of the cervix was the resultant outcome.

Brandt *et al*. (2019) compare the oncology and perioperative outcomes in patients who underwent minimally invasive surgery (MIS) to those who had laparotomy for newly diagnosed early-stage cervical carcinoma. Results revealed the rate of post-operative complications was notably lower in the MIS group than in laparotomy group. Survival patterns and treatment outcomes in patients with head and neck cancer (HNC) was assessed by (Okwor *et al*., 2017). It was concluded that high prevalence of the disease in men and patients that presented at early stages had higher survival than advance stage presentation. Also patients that received combined chemotherapy and radiotherapy had higher survival compared to those who had a single modality of treatment.

Furthermore, (Ren *et al*., 2018) conducted a study to investigate the racial disparities in the presentation, treatment and survival time of patients with hepatocellular carcinoma (HCC) or intrahepatic cholangiocarcinoma (ICC) between Chinese and other racial groups. Result showed that race was an important independent predictive factor in the HCC group while in the ICC group it was not important. In another study conducted by (Sun *et al*., 2018) to evaluate the effect of adding radioactive iodine (RAI) therapy to total thyroidectomy (TT) on overall survival (OS) in patients diagnosed with papillary thyroid cancer (PTC) cervical pathologically proves lymph node (LN) metastases (PNI). Findings showed that patients treated with TT + RAI had significant improvement in OS compared to those treated with only TT. The outcome of survival among patients with invasive lobular carcinoma (ILC) to ascertain the potential benefit of contralateral prophylactic mastectomy (CPM) was studied by (Yu *et al*., 2018). The result revealed that CPM does not offer any survival advantage to patients with invasive lobular carcinoma (ILC). Chen *et al*. (2021) evaluate the independent predictive rate of clinical and possible predictive factors in progression-free survival (PFS) in cervical cancer. The result showed the possible predictive factors for cervical cancer patients were number of lymph nodes, age at onset of symptoms, uterine manipulator and retrieved lymph nodes count combining with FIGO staging.

2.10 Research Gap

Many works reviewed assess the risk factors and associated factors of death using semi-parametric model. A few studies that determine the prognostic factors of cervical cancer to my knowledge use cox proportional hazard model which assume that the hazard does not follows any statistical distribution and also measure only the hazard rather than survival time are Gurmu, (2018), who used cox proportional hazard model to determine the risk factors affecting survival of cervical cancer, Vishma *et al*., (2016), Anfinan & Sait, (2020) and Marscarello *et al.*, (2013) all use cox proportional hazard model to identify the prognostic factors. However, in all this studies none was in Nigeria and none employ the use of parametric model. Consequently, this study intends to close the gaps by using parametric model to determine the prognostic factors associated with cervical cancer.

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1 Data Source

The data used for this research work is a secondary data obtained from reviewed case folders of all women diagnosed with cervical cancer, treated and followed up at the Oncology Department of the National Hospital Abuja (NHA), Nigeria, between January 2010 and December 2020. A total of 689 case folders were retrieved from the hospital libraries and patients' information were sorted and documented. The dataset has an identification for each patient, reporting date, summary, location, tribe, gender, age, hospital visit, death and recovered. Only 388 subjects (patients) fulfilled inclusion criteria (patients with complete information) and were included for further studies. All patients with missing follow-up data were excluded. Approval for the data used was granted by the hospital management.

3.2 Data Collection

Data collected for analysis include baseline demographic data (age, marital status, religion, tribe, occupation, parity), time variables including age at diagnosis, age at first birth, age at last birth, outcome data including events occurring during the follow-up period (recurrence, death) and 11-year status (alive with/without disease, deceased, censored), management data including radiotherapy, chemotherapy, combination of radiotherapy and chemotherapy and neither of the two, tumour characteristics including FIGO stage, tumour grade and histological type, comorbidity type (HIV, hypertension, diabetes and asthma.), family history, smoking status, menopause, alcohol consumption, menarche, coitarche and comorbidity.

3.3 Study Variables

Time to death (measured in month) of women with cervical cancer was used as the response variable in this study. Sociodemographic and socioeconomic factors (age, marital status, religion, tribe, occupation, smoking status, alcohol consumption), reproductive factors (menarche, coitarche, parity, age at menopause, age at first birth, age at last birth), pathological and clinical factors (FIGO stage, histology type, recurrence, tumour grade, family history, comorbidity, types of comorbidity) and treatment related factors (radiotherapy, chemotherapy, combination of the two, none) were independent variables considered in the study. All cause death of patients was the event of interest. All patients lost to follow-up and those who did not experience the event up to the end of study were censored.

3.4 Data Processing and Analysis

3.4.1 Descriptive statistics

Descriptive statistics was performed on the data. Quantitative (continuous) variables were presented as mean, median, standard error (SE), skewness and kurtosis, while qualitative (categorical) variables were presented as counts (frequency) and percentage. Patients were grouped base on occupation into civil servants (C/S), business, farmers, house wife (H/W) and others. Treatment taken include radiotherapy, chemotherapy, combination of the two (radio/chemo) and none. Cancer stage was grouped into Stages IA, IB, IIA, IIB, IIIA, IIIB, IVA and IVB according to International Federation of Gynaecology and Obstetrics (FIGO) staging for carcinoma of the vulva, cervix and endometrium. Subjects were grouped as married, widowed, divorced, separated and single according to their marital status. Islam and Christianity were the groups used for religious categorisation. Subjects were classified into the three major language of Nigeria- Hausa, Igbo, Yoruba and all subjects that falls outside these three were grouped as others. Subjects were classified either smokers and or non-smokers base on their smoking status. Categorical variables- family history, alcohol consumption, comorbidity and recurrence were coded Yes or No. Histological types of the subjects were squamous cell carcinoma (SCC), adenocarcinoma and adenosquamous. Tumor grade of the subjects were grouped into poorly differentiated (PD), well differentiated (WD) and moderately differentiated (MD).

3.4.2 Kaplan-Meier (K-M) analysis

Kaplan-Meier estimator of survival function was used to estimate the survival rate and the confidence interval method was used to estimate the median survival time of the patients, as the median is less affected by outliers and skewed data. Kaplan-Meier curve was estimated for qualitative variables in the study. The idea of this method is based on the probability of surviving in k or more periods in the study and is a product of k probabilities when each period is observed under it (Bewick *et al*., 2004). It is express mathematically by equation (3.1).

$$S(k) = P_1 \times P_2 \times P_3 \times \ldots \times P_k \ldots\ldots\ldots \qquad (3.1)$$

Where

P_1 constitutes surviving proportion in the first period, P_2 is the proportion survived over the second period and P_k the proportion survived over period k.

3.4.2.1 Mathematical formulation of Kaplan-Meier model

The Kaplan-Meier was calculated for each variable by taking the product of proportion of patients at risk at that time minus number of deaths divided by number at risk. The Kaplan-Meier survival estimator is given by equation (3.2).

$$S_{KM}(t) = \prod_{1:\, t_i < t} \frac{n_i - d_i}{n_i} \tag{3.2}$$

Where

t_i = time point

n_i = number at risk at time t

d_i = number of death at time t_i (Etikan *et al*., 2017)

3.4.2.2 Assumptions of K-M estimator

(i) The data is composed of two mutually exclusive and exhaustive state known as event or censored

(ii) The survival time should be clearly defined and accurately measured

(iii) The data is right censored

(iv) The event and censoring are independent

(v) No trend is observed in the data

(vi) Right censoring is similar in all the groups (variables) (Adamu *et al*., 2019).

The six assumptions were checked before the analysis was performed and non was violated.

3.4.2.3 Algorithm of K-M estimator

The raw data was stored in MS Excel format using actual calendar date and time. Serial time was calculated automatically and was used in curve construction and data analysis.

Table construction was carried out on Excel spread sheet containing three key elements required for input:

(i) Serial time (time from diagnosis of cervical cancer till death (event) or alive (censored) patients)

(ii) Status at serial time (0 = alive, 1 = death)

(iii) The study group (the variable being estimated). 95% confidence interval (CI) was used

The survival curve was plotted with cumulative probability (Y) axis measured in percentage against time (X) axis measured in months.

3.4.2.4 Computation of 95% confidence interval (CI)

The 95% confidence interval for Kaplan-Meier (K-M) curve was computed using the relationship

$$S_{KM}(t) \pm 1.96\sqrt{\mathrm{var}[S_{KM}(t)]}$$

Where Greenwood's formula var[S_{KM}(t)] is given by equation (3.3).

$$\mathrm{var}[S_{KM}(t)] = (S_{KM}(t))^2 \sum_{i\,:\,t_{(i)}\leq t} \left[\frac{m_i}{n_i(n_i - m_i)}\right] \tag{3.3}$$

Where

t_i is the i-ordered failure time, m_i is the number of failures at $t_{(i)}$ and n_i is the number in the risk set at $t_{(i)}$ (Kleinbaum and Klein, 2012).

S_{KM} (t) is the estimated survival probability from Kaplan-Meier curve at the true median survival time.

3.4.3 The Log-Rank test

The difference in survival time among different groups were compared using Log-rank test with the chi square test statistic.

$$\chi^2 = \sum \frac{\left(\sum O_{jt} - \sum E_{jt}\right)^2}{\sum E_{jt}} \tag{3.4}$$

Where

J = 1, 2, . . ., n

ΣO_{jt} is the sum of the observed number of events in the jth group over time (t)

ΣE_{jt} is the sum of the expected number of events in the jth group over time (t)

With k – 1 degrees of freedom (df) where k represents the number of comparison groups

3.4.3.1 Assumptions of the Log-Rank test

(i) Censored and the uncensored patients have the same probability of the event (censoring is non-informative)

(ii) Kaplan-Meier curves of the group did not intersect

(iii) No particular distribution for the survival curve is assumed (distribution free) (Emmert-Streib and Dehmer, 2019).

3.4.3.2 Algorithm for the Log-Rank test

(i) The hypothesis was set up and the level of significance determined

H_o: there is no difference in survival between the groups

H_1: There is a difference in survival between the groups (P-value of <0.05)

(ii) Select the appropriate test statistics. The test statistic for this study is given by equation (3.5).

$$\chi^2 = \sum \frac{\left(\sum O_{jt} - \sum E_{jt}\right)^2}{\sum E_{jt}} \tag{3.5}$$

Where

ΣO_{jt} is the sum of the observed number of events in the jth group over time (t)

ΣE_{jt} is the sum of the expected number of events in the j^{th} group over time (t), with k-1 degrees of freedom (df) where k represents the number of comparison groups. The expected deaths at time t was computed using equation (3.6).

$$e_{jt} = \frac{n_{jt}}{\prod_{u=1}^{\infty} n_{jt}} \times d_i \tag{3.6}$$

Where

di is the total deaths in all groups at time t, n_{jt} is the total number of patients at risk in the j^{th} groups.

The total number of deaths expected in the groups is computed as

$$E_1 = \sum_{\forall t} e_{1t}, \ E_2 = \sum_{\forall t} e_{2t}, \ E_3 = \sum_{\forall t} e_{3t} \ldots E_j = \sum_{\forall t} e_{jt} \tag{3.7}$$

Thus the log-rank test statistic (LRstat) was computed as

$$LR_{stat} = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \tag{3.8}$$

For i = 1, 2, 3, . . ., n, with n being the number of patients.

(iii) Setting up the decision rule. The decision rule is to reject H_o if p-value < α value (0.05) or fail to reject when p – value is large.

(iv) Compute the test statistic. The test statistic was done using R statistical package

(v) Conclusion will depend on the outcome of step (iv).

3.4.4 Classification tree

The following notations are used in defining the classification tree model.

$\Pi_{i,}$ i = 1, 2, . . . , C prior probability of each classes

L (i, j) i, j = 1, 2, . . ., C loss matrix for misclassifying class i as class j

A = some node in the tree

P(A) = probability for future observations can be classified in the node A, P(A_L), P(A_R) denote the left and right node son under parent A

I(A) = impurity measurement of node A

N_i(A) = number of observations of class "i" in node A

N_i = number of observations of class "i" in the whole learning data set

R(A) = risk of node A

3.4.4.1 Splitting criterion and impurity measurement

Let C be the classes considered in the classification, the classification tree is grown under the splitting criterion that minimises the impurity of the nodes in the tree. To achieve this, impurity measurement function f were introduced. (P_{iA}) denotes the impurity in the node A caused by class i. Intuitively and most commonly, we need the node with P_{iA}, estimated by the frequency of class i in node A to be as far from 1/C as possible. Gini index used to identify the best split is defined by equation (3.9).

$$P_{iA} = P(1 - P) \tag{3.9}$$

Where

P is the relative frequency of misclassification.

To summarise the total impurity of node A, we sum the impurity measurement of each class in node A using equation (3.10).

$$I(A) = \sum_{i=1}^{C} f(P_{iA}) \tag{3.10}$$

All possible splitter variables with all possible splitting values are first calculated for the node A. The best splitter was selected so that the average impurity reduction by two son nodes is maximised.

$$\mathrm{\Delta I(A) = P(A)I(A) - P(A_L)I(A_L) - P(A_R)I(A_R)} \quad (3.11)$$

$$\mathrm{\{A_L, A_R\} = argmax\Delta I(A)} \quad (3.12)$$

The branches of the tree will continue splitting until either of the two following conditions are met.

i. The number of the observations in terminal node reaches the minimum predefined (20 in our class)
ii. All the observations in the terminal node have same value for every predictor.

3.4.4.2 Receiver operating characteristic (ROC) curve

This section introduces criteria to evaluate model prediction. Receiver operating characteristic (ROC) curves plot sensitivity by 1- specificity of a binary classifier across different thresholds. Sensitivity and specificity are defined by equation (3.13) and (3.14) respectively.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3.13)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (3.14)$$

ROC curves are generated by plotting sensitivity against 1-specificity. It can be shown that the area under the curve (AUC) is the estimate of the probability of the classifier to rank a randomly chosen positive event higher than a randomly chosen negative event using normalised unit.

3.4.4.3 Variable importance

The Gini index was used to measure the importance of variable. For a given variable, the importance is calculated by summing the Gini index across all the nodes which were split by the variable. Two facts should be noticed:

i. The summation was weighted by the position of the node: a more ancestor nodes receive more "credit" in importance as more data points were split by it.

ii. Using the standard 0 and 1 cost, the Gini index equals to the decrease in the misclassification rate.

3.4.4.4 Confusion matrix

The confusion matrix was used as a tool for predictive analysis in the classification and regression tree (CART). It was used to check the performance of the model. The data was divided into training and testing sets.

Table 3.1: 2×2 confusion matrix for CART model

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Where

True positives (TP): The actual value was positive and the model predicted a positive value.

False positives (FP): The prediction is positive and it is false (also known as type I error).

False negatives (FN): The prediction is negative and the result is also false (also known as type II error).

True negatives (TN): The actual value was negative and the model predicted a negative value.

3.4.5 Accelerated failure time model

The Cox proportional hazard assumptions was tested on the data and found to have failed, hence, the choice of accelerated failure time (AFT) model. The AFT model was used to identify the prognostic factors associated with the 388 cervical cancer patients. The exponential, Weibull, log-normal and log-logistic were the AFT models considered. The models were compared using statistical criteria-maximum likelihood (ML) test, Akaike information criteria (AIC) and Bayesian information criteria (BIC) to find the best fitted model for the data.

3.4.5.1 Mathematical formulation of AFT

Let T be a random variable of survival times and X is a column vector of the covariates $X_1, X_2, \ldots, X_p$, the AFT model defines the relationship between covariates and the survival time as a linear relation between natural logarithm of survival time and the covariates X. This relationship is given by equation (3.15).

$$Y = \ln T = \mu + \theta^t X + \sigma W \tag{3.15}$$

Where

μ is the slope,

$\sigma > 0$ is an unknown scale parameter

$\theta^t = (\theta_1, \theta_2, \ldots, \theta_p)$ is a vector of regression coefficients

$\theta = -\beta$

σ is a scale parameter

W is a distribution error which is a random variable and assumed to follow a certain parametric distribution.

For every distribution of W, there is a related parametric for T. The name for the AFT model comes from the distribution of T rather than the parametric distribution of lnT (Alfensi, 2018).

The survival function of T_i, i = 1, 2, . . ., n is given by equation (3.16).

$$\begin{aligned} S_i(t) &= \Pr(T_i \geq t) \qquad (3.16) \\ &= \Pr(\ln T_i \geq \ln t) \\ &= \Pr(Y_i \geq \ln t) \\ &= \Pr(\mu + \theta^t X + \sigma W \geq \ln t) \\ &= \Pr\left(W_i \geq \frac{\ln t - (\mu + \theta^t X)}{\sigma}\right) \end{aligned}$$

According to the AFT process, it is assumed that the influence of the covariates has an effect on the logarithmic time scale and therefore, multiplicatively on the time scale itself (Ponnuraja and Venkatesan, 2010). The assumption is expressed by equation (3.17).

$$S(t|X) = S_o[t \exp(\beta^t X)] \qquad (3.17)$$

Where

S(t|X) is the survival function at time t

S_o is the baseline survival function at the time t

$\beta^t = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients, and n ϵ N.

The factor exp($\beta^t X$) in equation (3.17) is known as the acceleration factor on the time scale of t, which accelerates the survival function with covariate X = 0 (Alfensi, 2018). The acceleration factor is the key measure of association obtained in the AFT model (Saikia and Barman, 2017).

3.4.5.2 Weibull AFT model

The survival function of survival time T and the hazard function for Weibull distribution with parameters λ and α is given by equation (3.18) and (3.19) respectively.

$$S(t) = \exp(-\lambda t^{\alpha}) \tag{3.18}$$

$$h(t) = \lambda\alpha t^{\alpha-1} \tag{3.19}$$

if X in equation (3.15) equal to zero, the equation becomes

$$Y = \mu + \sigma W \tag{3.20}$$

Where

$\mu = (-\ln(\lambda)/\alpha)$, $\sigma = \alpha^{-1}$ and W follows the standard extreme value distribution with probability function given by equation (3.21).

$$f(w) = e^{(w-e^{w})} \tag{3.21}$$

Integrating the probability density function in equation (3.21) with respect to variable u on the interval from w and ∞, the survival function of W is given by equation (3.22).

$$S(W) = \int_{w}^{\infty} f(u)du \tag{3.22}$$

$$= \int_{w}^{\infty} e^{(u-e^{u})}\,du$$

$$= e^{(-e^{w})}$$

The hazard function of W is given by equation (3.23).

$$h(w) = \frac{f(w)}{S(w)} \tag{3.23}$$

$$= \frac{e^{(w-e^{w})}}{e^{(-e^{w})}}$$

$$= e^{w}$$

Therefore, the survival function of multivariate Weibull AFT model is expressed by equation (3.24).

$$S(y|X) = \exp\left[-\exp\left(\frac{y-(\mu+\theta^{t}X)}{\sigma}\right)\right] \tag{3.24}$$

3.4.5.3 Exponential AFT model

The exponential AFT model can be derived from Weibull distribution by taking σ = 1 or α = 1 so that equation (3.24) becomes equation (3.25).

$$S(y|X) = \exp\left[-\exp\left(y-(\mu+\theta^{t}X)\right)\right] \tag{3.25}$$

Where

$S(y|X)$ is the survival function for exponential AFT model in multivariate case.

3.4.5.4 Log-normal AFT model

If the random variable of the survival time T is assumed to follow a log-normal distribution, the baseline survival function $S_o(t)$ and baseline hazard function $h_o(t)$ are expressed by equation (3.26) and (3.27) respectively (Alfensi, 2018).

$$S_{o}(t) = 1-\varphi\left(\frac{\ln t-\mu}{\sigma}\right) \tag{3.26}$$

$$h_o(t) = \frac{\varphi\left(\frac{\ln t}{\sigma}\right)}{1 - \Phi\left(\frac{\ln t}{\sigma}\right)} \tag{3.27}$$

Where

μ and σ are intercept and scale parameters respectively.

φ(x) is density probability function at time x

Φ(x) is the cumulative distribution function at time x

Therefore, for a given set of covariates $X = (X_1, X_2, \ldots, X_p)^t$ the survival function is represented by equation (3.28).

$$S(t) = S_o[t \exp(\beta^t X)] \tag{3.28}$$

$$= 1 - \varphi\left(\frac{(\ln t - (\mu + \theta' X))}{\sigma}\right)$$

Where

t ϵ T is survival time

$\beta^t = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector coefficients.

3.4.5.5 Log-logistic AFT model

If the ε_i has the logistic distribution then T_i follows the log-logistic distribution. The survival function of logistic distribution according to (Saikia and Barman, 2017) is given by equation (3.29).

$$S_{\varepsilon_i}(\varepsilon) = \frac{1}{1 + e^{\varepsilon}} \tag{3.29}$$

The survival function of log-logistic AFT model is represented by equation (3.30).

$$S_i(t) = \left\{ \frac{1}{1 + e\left(\frac{\log t - \mu - \beta_1 X_1 - \cdots - \beta_p X_p}{\sigma}\right)} \right\} \tag{3.30}$$

3.4.5.6 Parameter estimation of accelerated failure time model

AFT models are fitted by using maximum likelihood estimation (MLE) method. The likelihood of an n observed survival times $t_1, t_2, \ldots, t_n$ with known parameters $\beta^t = (\beta_1, \beta_2, \ldots, \beta_p)$, μ and σ which contain (n-r) right censored data for $0 \leq r \leq n$ (Alfensi, 2018) is given by equation (3.31) and the event indicator for the ith individual δ_i is given by equation (3.32).

$$L = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i} \tag{3.31}$$

Where

$f_i(t_i)$ and $S_i(t_i)$ are the density function and survival function for the ith individual at time t_i respectively.

$$\delta_i = \begin{cases} 1, & \text{if the i}^{\text{th}} \text{ observation is event} \\ 0, & \text{if the i}^{\text{th}} \text{ observation is censored} \end{cases} \tag{3.32}$$

Now

$$S_i(t_i) = S_{\varepsilon_i}(W_i)$$

$$f_i(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}(W_i) \tag{3.33}$$

The log-likelihood function is given by equation (3.34).

$$\log L = \sum_{i=1}^{n} (\sigma t_i)^{-\delta_i} \left\{f_{\varepsilon_i}(W_i)\right\}^{\delta_i} \left\{S_{\varepsilon_i}(W_i)\right\}^{1-\delta_i} \tag{3.34}$$

$$\log L = \sum_{i=1}^{n} -\delta_i \log(\sigma t_i) + \delta_i \log\{f_{\varepsilon_i}(W_i)\} + (1-\delta_i)\log\{S_{\varepsilon_i}(W_i)\} \qquad (3.35)$$

Where

$W_i = \log T - \mu - \beta_1 x_1 - \beta_2 - \cdots - \beta_p x_p$ and MLE of $(P+2)$ unknown parameters μ, σ and $\beta_1, \ldots, \beta_p$ are found by maximising the log-likelihood function using Newton Raphson procedure (Saikia and Barman, 2017).

3.4.5.7 Model selection

To check the appropriate AFT model for the analysis of the data, Akaike information criteria (AIC) and Bayesian information criteria (BIC) were used. AIC was computed using equation (3.36) while BIC was computed using equation (3.37).

$$\mathrm{AIC} = -2\mathrm{LL} + 2(P + C) \qquad (3.36)$$

$$\mathrm{BIC} = -2(\mathrm{LL}) + (P + C) \times \log(n) \qquad (3.37)$$

Where

P is the number of parameters in the distribution

C is the number of coefficients (excluding constant) in the model.

n is the number of observations

LL (log-likelihood) is the logarithm of the similarities of the model.

P = 1 for exponential, P = 2 for Weibull, Log-normal and Log-logistic. The model with the smallest AIC and BIC value was considered as the best fitted model.

CHAPTER FOUR

4.0 RESULTS AND DISCUSSION

4.1 Descriptive Statistics

Table 4.1 shows the baseline characteristics for quantitative variables. Both the average (mean) age and the median age of the patients is 54 years with standard error (SE) of 0.62. Twenty three (23) and ninety (90) years were the minimum and maximum ages of the patients. The skewness of 0.22 shows the age distribution is approximately symmetric while a kurtosis of -0.37 is an evidence that the age is light tailed.

The total number (N) of patients' age, age at first birth, age at last birth, menarche, coitarche and parity is 388. Nineteen (19) years is the mean age at first birth with the maximum age being 41 years. While the mean age at last birth is 37 years with maximum being 53 years. Looking at the skewness, age at first birth is highly skewed while age at last birth is moderately skewed. Kurtosis of 3.84 of age at first birth shows that it is heavily tailed while that of age at last birth that has a value of 0.11 indicate that the distribution is approximately normal. The mean menarche is age 14 years with the maximum age being 20 years and has a skewness of 0.17. This is an evidence of symmetric distribution with a kurtosis of 0.18 which is approximately normally distributed. The mean age of the first sexual interaction of a girl with sex partner (coitarche) is 17 years with the maximum of 31 years. The skewness value of 1.09 for coitarche implied that the distribution is highly skewed and has a kurtosis of 2.14 which is an indication that it is heavily tailed distribution. Furthermore, the mean parity is 6 years and the maximum is 14 years, while the mean menopause recorded is 49 years and the maximum is 62 years with a skewness value of 0.09 which shows that it is fairly symmetric and kurtosis of -0.45 indicate a light tailed distribution.

Table 4.1: Baseline characteristics of 388 cervical cancer patients for quantitative variables

Variables	N	Mean	Std Error	Min	Med	Max	Skewness	Kurtosis
Age	388	54.43	0.62	29	54	90	0.22	-0.37
Age at first birth	388	19.12	0.19	5	18	41	1.20	3.84
Age at last birth	388	36.53	0.34	15	37	53	-0.41	0.11
Menarche	388	14.12	0.08	10	14	20	0.17	0.18
Coitarche	388	17.18	0.14	12	17	31	1.09	2.14
Parity	388	6.00	0.13	1	6	14	0.19	-0.42
Menopause	273	49.49	0.33	35	50	62	0.09	-0.45

Table 4.2 shows the baseline characteristics for qualitative variables. 17(4.38%) of the patients were divorced, 248(63.92%) were married, 16(4.12%) were separated, 10(2.58%) were single and 97 (25.00%) were widowed. Islam and Christianity has 145 (37.37%) and 243 (62.63%) respectively. FIGO stage showed stage IA 2 (0.52%), IB 30 (7.73%), IIA 44 (11.34%), IIB 101 (26.03%), IIIA 63 (16.24%), IIIB 102 (26.03%), IVA 33 (8.51%), IVB 13 (3.35%). A group of 353 (90.98%) of the patients had no recurrence, while 35 (9.02%) patients experienced recurrence of cervical cancer. The family history showed that 82.99% had no history of cancer while 17.01% had history of cancer in the family. Data revealed that 28 (16.44%) of the patients were smokers or had history of smoking while 363 (93.56%) never smoked. Tribe showed Hausa (15.21%), Yoruba (10.31%), Igbo (14.69%) and 59.79% represents other ethnic nationalities. The most frequent histological type was squamous cell carcinoma SCC (n=328, 84.54%), followed by adenocarcinoma (n=46, 11.86%) and adenosquamous (n=14, 3.61%). The tumour

grade showed poorly differentiated 41.24% to be the most common among the patients followed by moderately differentiated 32.47% and well differentiated 26.29%. Patients that received chemo/radiotherapy 148 (38.14%), chemotherapy 72 (18.56%), radiotherapy 103 (26.55%) and none 65 (16.75%). Patients with comorbidity were 146 (37.63%) while those that had no comorbidity were 242 (62.37%).

Table 4.2: Baseline characteristics of 388 cervical cancer patients for qualitative variables

Variable	Category	Count	Percentage (%)
Marital Status	Married	248	63.92
	Widowed	97	25.00
	Divorced	17	4.38
	Separated	16	4.12
	Single	10	2.58
Religion	Islam	145	37.37
	Christianity	243	62.63
FIGO Stage	Stage I A	2	0.52
	Stage I B	30	7.73
	Stage II A	44	11.34
	Stage II B	101	26.03
	Stage III A	63	16.24
	Stage III B	102	26.29

Table 4.2: Continued

	Stage IV A	33	8.51
	Stage IV B	13	3.35
Occupation	Civil Servant	53	13.66
	Business	134	34.54
	Farmer	33	8.51
	House wife	136	35.05
	Others	32	8.25
Recurrence	Yes	35	9.02
	No	353	90.98
Family history	Yes	66	17.01
	No	363	82.99
Smoking Status	Yes	28	6.44
	No	363	93.56
Alcohol consumption	Yes	89	22.94
	No	299	77.06
Tribe	Hausa	59	15.21
	Yoruba	40	10.31
	Igbo	57	14.69
	Others	232	59.79

Table 4.2: Continued

Histology	Squamous cell carcinoma	328	84.54
	Adenocarcinoma	46	11.86
	Adenosquamous	14	3.61
Tumor grade	Well differentiated	102	26.29
	Moderately differentiated	126	32.47
	Poorly differentiated	160	41.24
Treatment taken	Radiotherapy	103	26.55
	Chemotherapy	72	18.56
	Radio/Chemo	148	38.14
	None	65	16.75
Comorbidity	Yes	146	37.63
	No	242	62.37

4.2 Kaplan-Meier (K-M) Analysis

4.2.1 Overall K-M analysis for cervical cancer patients

Figure 4.1 and Table 4.3 shows that out of 388 cervical cancer patients 234 deaths (60% of number of patients) were recorded. Each patients has a 50 % chance of surviving at least 13 months and a minimum of 10 months but not more than 17 months when the cancer stage and other covariates are not put into consideration.

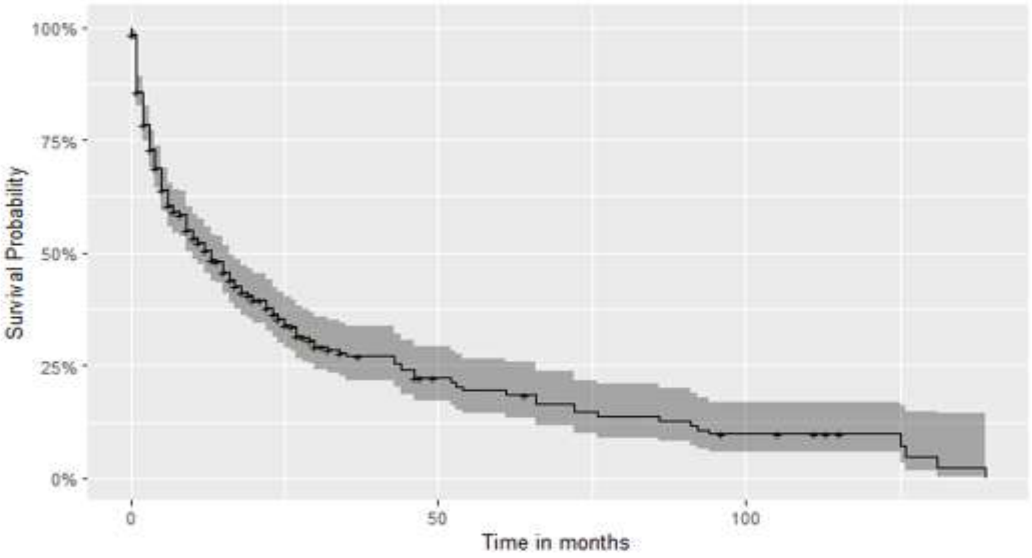


Figure 4.1: Overall survival curve for cervical cancer patients

Table 4.3: K-M estimate of the overall survival function

No. of Patients	Events	Median	Std. Error	0.95LCL	0.95UCL
388	234	13	1.79	10	17

4.2.2 Estimate of survival base on marital status

Figure 4.2 and Table 4.4 shows distinct survival time between the categories of status of marriage. Divorced patients have a total number of 17 out of which 9 (53%) experienced the event of interest (death) with median survival time of 12 months. Out of 248 number of married patients, 149 (60%) experienced the event of interest with median survival time of 13 months. Separated patients have a total of 16 out which 9 (56%) was event with 24 months of median survival time. Single patients have a total of 10 out of which 8 (80%) being the number of events with median survival time of 8 months. Widowed patients with total number of 97 had 59 (61%) as event with 9 months median survival time. It was observed that single patients experienced more deaths while divorced patients experienced least deaths. However, median survival time for separated patients (24 months) is higher than those in other categories with widowed patients showing the least median survival time of 9 months.

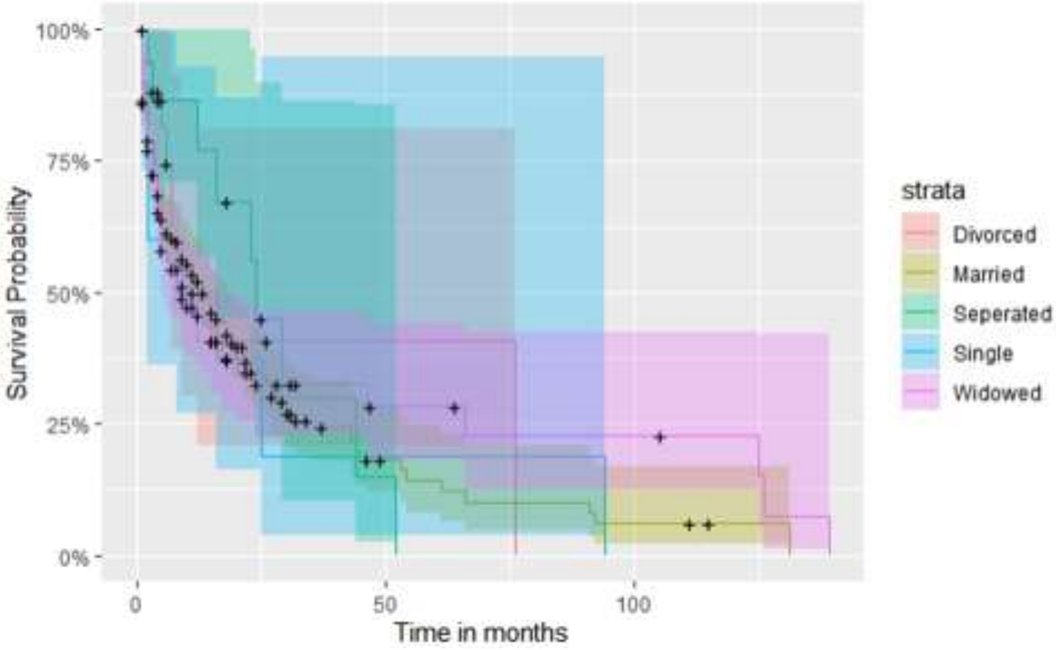


Figure 4.2: Estimate of survival curve base on marital status

Table 4.4: K-M estimate of survival time base on marital status

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Divorced	17	9	53	12	3.79	4.57	19.43
Married	248	149	60	13	2.22	8.66	17.35
Separated	16	9	56	24	1.42	21.22	26.78
Single	10	8	80	8	9.84	0.00	27.28
Widowed	97	59	61	9	3.29	2.55	15.45

4.2.3 Estimate of survival base on religion

Figure 4.3 and Table 4.5 displays the religion of the patients. Christianity has a total number of 243 patients of which 141 (58%) were number of events with a median survival time of 13 months while Islam has a total number of 145 patients out of which 93 (64%) were number of events with 13 months as the median survival time. It showed almost the same median survival time between the two religions, hence, no sign of difference in length of survival based on religion.

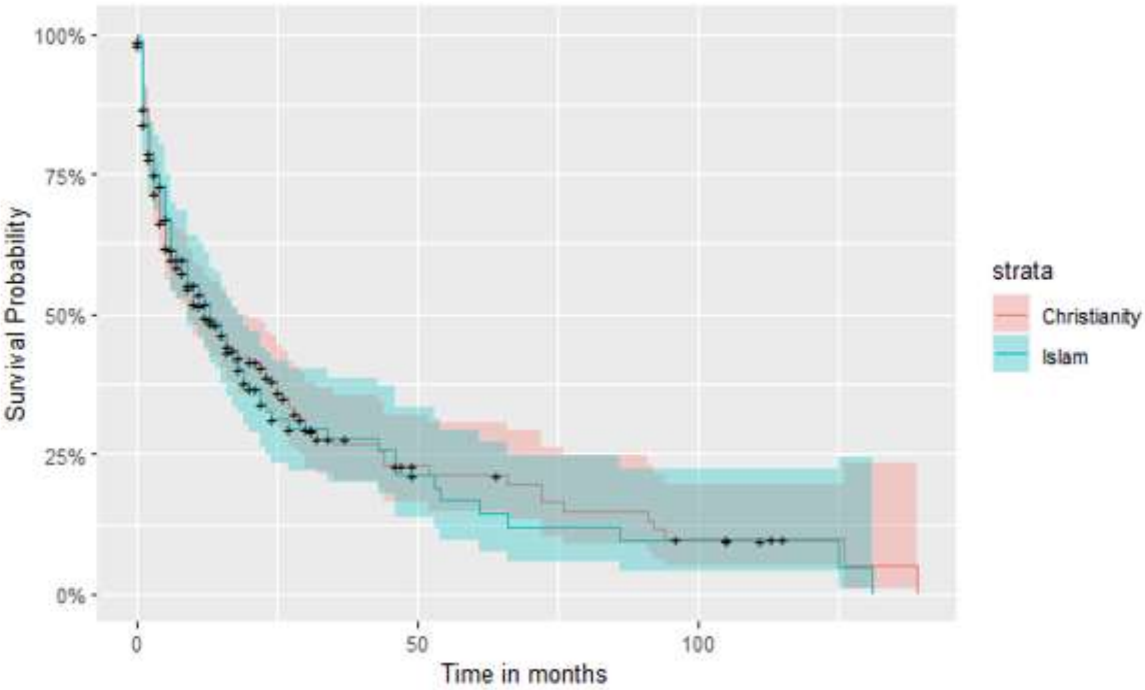


Figure 4.3: Estimate of survival curve base on religion

Table 4.5: K-M estimate of survival time base on religion

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Christianity	243	141	58	13	2.75	7.62	18.38
Islam	145	93	64	13	2.49	8.13	17.88

4.2.4 Estimate of survival base on tumour stages

Figure 4.4 and Table 4.6 showed a distinct survival time between the stages of the tumour. Patients who were diagnosed at an early stage had a better survival compared to those who were diagnosed at a later stage. Stage IA has a total number of 2 patients of which 1

(50%) was event with a median survival time of 76 months. Stage IB has a total number of 30 patients out of which 10 (33%) was event with 66 months as the median survival time. Stage IIA and IIB had a total number of 44 and 101 patients respectively out of which 14 (32%) and 58 (57%) were events with median survival time of 30 and 12 months respectively. Stage IIIA and IIIB had a total number of 63 and 102 patients, of which 44 (70%) and 81 (79%) were events with median survival time of 9 and 5 months respectively. Stage IV (A and B) had 33 and 13 patients in total out of which 18 (55%) and 8 (62%) with 15 and 2 months median survival time respectively.

A decreasing trend in the 50% chance of survival as the tumour stage increases was observed. Patients in stages IA tumour had highest median survival time while those in stage IVB tumour had the least survival time. However, it is unclear why patients in stage IVA have more median survival time (of 11 months) than those in stage IIIA and IIIB (9 and 5 months respectively), although more deaths were observed between stage IIIA and IIIB.

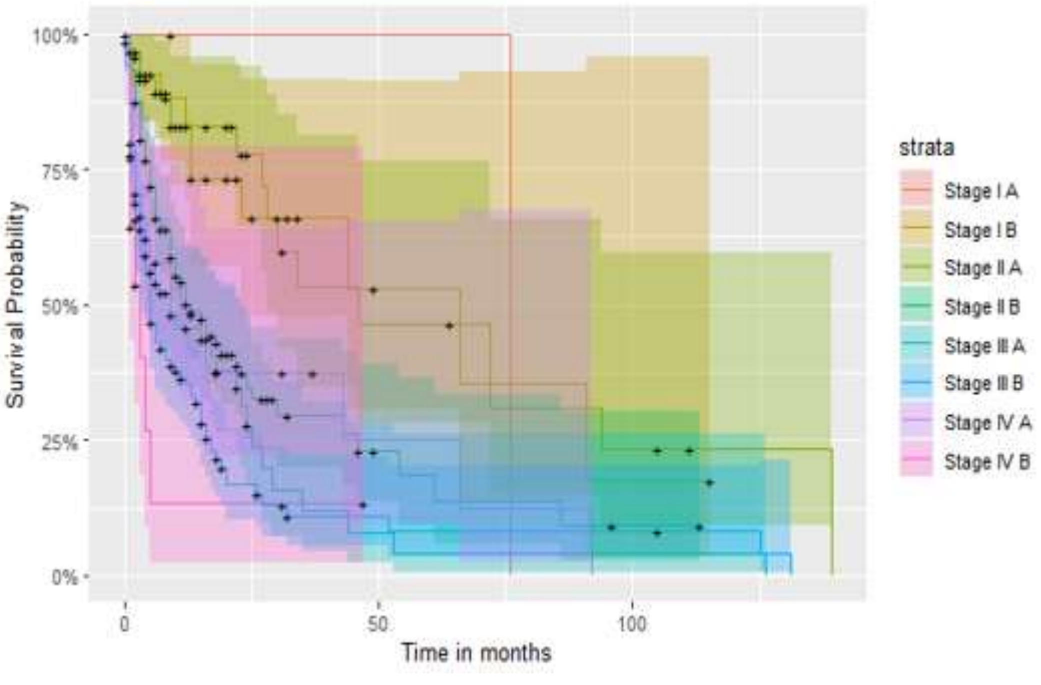


Figure 4.4: Estimate of survival curve base on tumour stages

Table 4.6: K-M estimate of survival time base on tumour stages

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Stage IA	2	1	50	76	-	NA	NA
Stage IB	30	10	33	66	24.29	18.39	113.61
Stage IIA	44	14	32	30	12.98	4.55	55.45
Stage IIB	101	58	57	12	2.98	6.17	17.83
Stage IIIA	63	44	70	9	5.31	0.00	19.40
Stage IIIB	102	81	79	5	0.74	3.55	6.45
Stage IVA	33	18	55	15	4.50	6.18	23.82
Stage IVB	13	8	62	2	1.07	0.00	4.11

4.2.5 Estimate of survival base on occupation

Figure 4.5 and Table 4.7 show a distinct survival time of the patients based on occupation. Patients with business as occupation have a total number of 134 out of which 74 (55%) been events with a median survival time of 13 months. C/S has a total number of 53 of which 32 (60%) been event with a median survival time of 11 months. Farmers had a total number of 33 out of which 16 (48%) been event with a median survival time of 25 months. House wives have a total number of 136 of which 96 (71%) been events with 9 months as the median survival time while others had a total of 32 out of which 16 (50%) been event with median survival time of 18 months. It was noted that farmers have a 50% chance of surviving at least 25 months while house wives have the least median survival time. Excluding others category, farmers' category also experienced a smaller number of deaths.

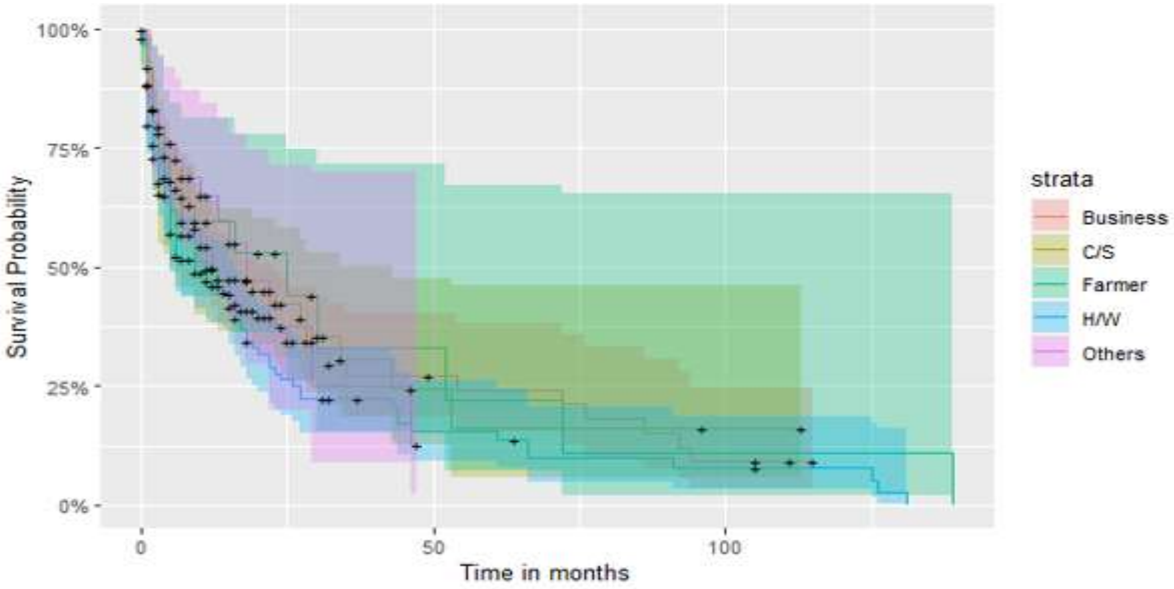


Figure 4.5: Estimate of survival curve base on occupation

Table 4.7: K-M estimate of survival time base on occupation

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Business	134	74	55	13	2.23	8.63	17.37
C/S	53	32	60	11	7.47	0.00	27.64
Farmers	33	16	48	25	12.93	0.00	50.34
H/W	136	96	71	9	3.22	2.69	15.31
Others	32	16	50	18	4.53	9.13	28.88

4.2.6 Estimate of survival base on recurrence

Figure 4.6 and Table 4.8 shows a slight difference in the proportion of deaths experienced after the cancer treatment. Recurrence occurred among 35 patients, and experienced 23 (66%) deaths against a total number of 353 that did not reoccur with 211 (60%) deaths. However, the median survival time is higher (43 months) for patients whose cancer reoccurred after treatment than those (12 months) who did not show a symptom of recurrence after treatment.

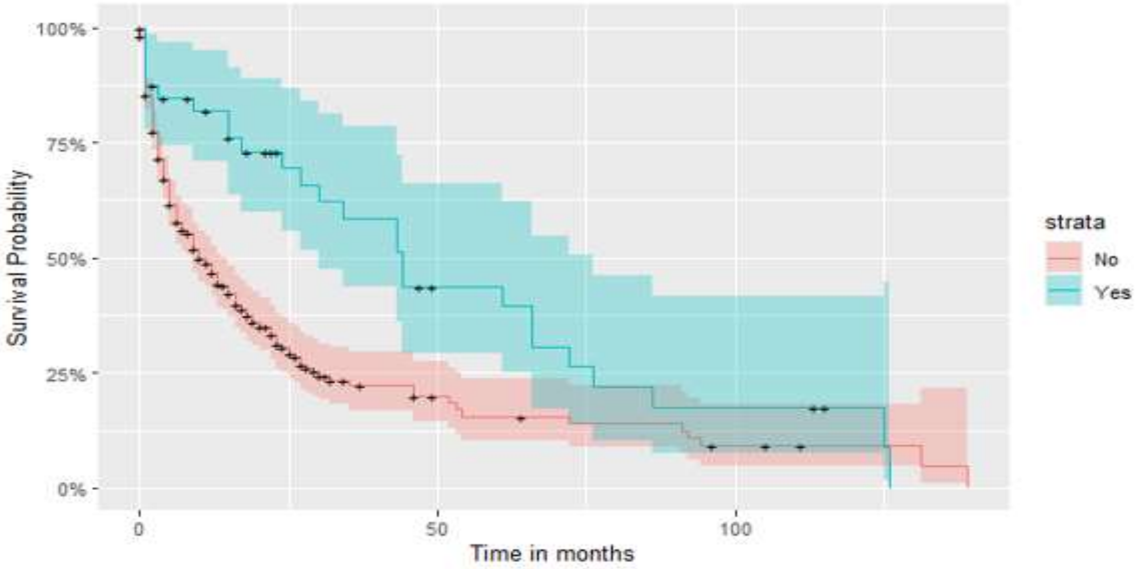


Figure 4.6: Estimate of survival curve base on recurrence

Table 4.8: K-M estimate of survival time base on recurrence

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Yes	35	23	66	43	7.92	27.48	58.53
No	353	211	60	12	1.51	9.04	14.96

4.2.7 Estimate of survival base on family history

Figure 4.7 and Table 4.9 displays patients that have family history of cervical cancer had a total number of 66 of which 37 (56%) died with a median survival time of 13 months while those without family history have a total number of 322 of which 197 (61%) been events with 13 months as the median survival time. It showed a trend of higher survival and less deaths among patients with family history of cervical cancer than those with no family history.

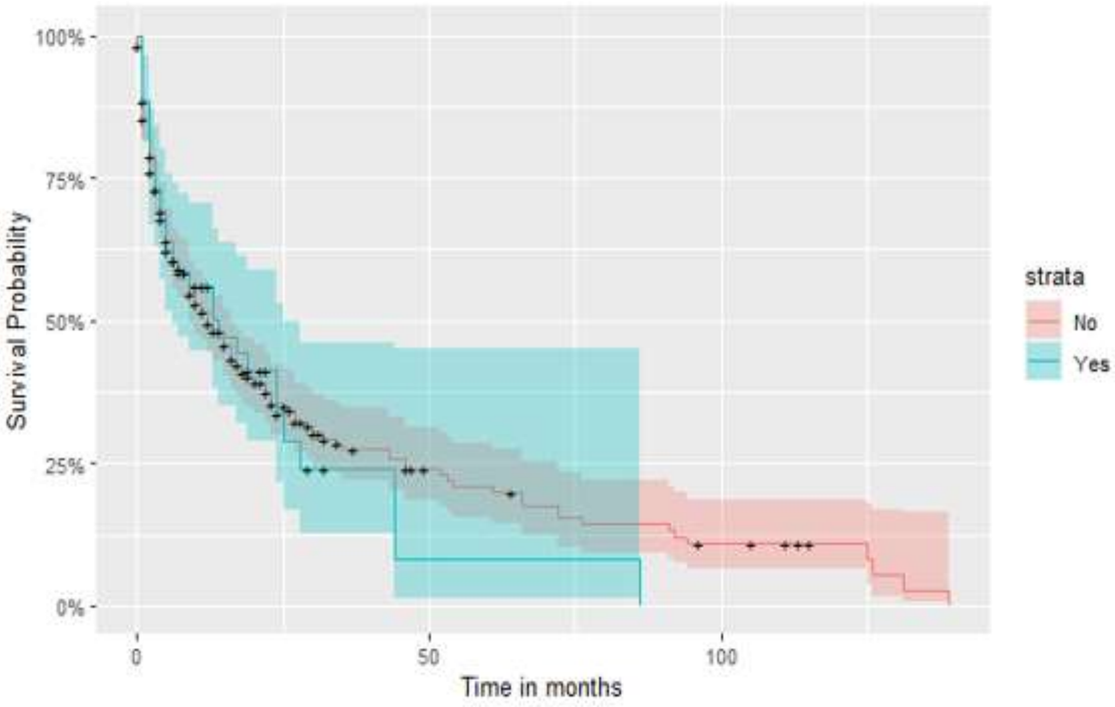


Figure 4.7: Estimate of survival curve base on family history

Table 4.9: K-M estimate of survival time base on family history

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Yes	66	37	56	13	4.68	3.83	22.17
No	322	197	61	13	1.89	9.31	16.69

4.2.8 Estimate of survival base on smoking status

Figure 4.8 and Table 4.10 displays smoking status. Both categories of patients had a total number of 25 and 363 of which 15 (60%) and 219 (60%) been event with 13 months of median survival time respectively. Patients who smoke or do not have similar median survival time of 13 months and a similar proportion of the patients in these categories experienced deaths.

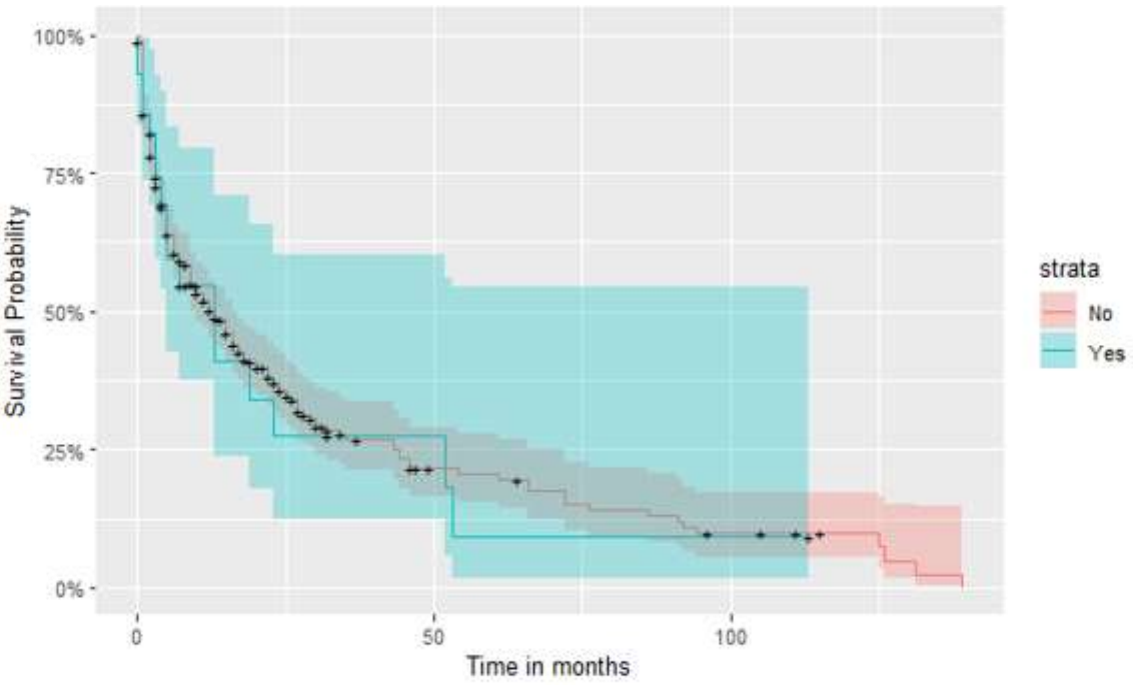


Figure 4.8: Estimate of survival curve base on smoking status

Table 4.10: K-M estimate of survival time base on smoking status

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Yes	25	15	60	13	4.65	3.89	22.11
No	363	219	60	13	1.88	9.31	16.69

4.2.9 Estimate of survival base on alcohol consumption

Out of 89 patients that consume alcohol 52 (58%) were deaths (events) with a median survival time of 13 months. Patients who don’t have a total number of 299 of which 182 (61%) (event) were deaths with a median survival time of 14 months. It was deduced that both categories have similar median survival time and similar proportion experienced death (Figure 4.9 and Table 4.11).

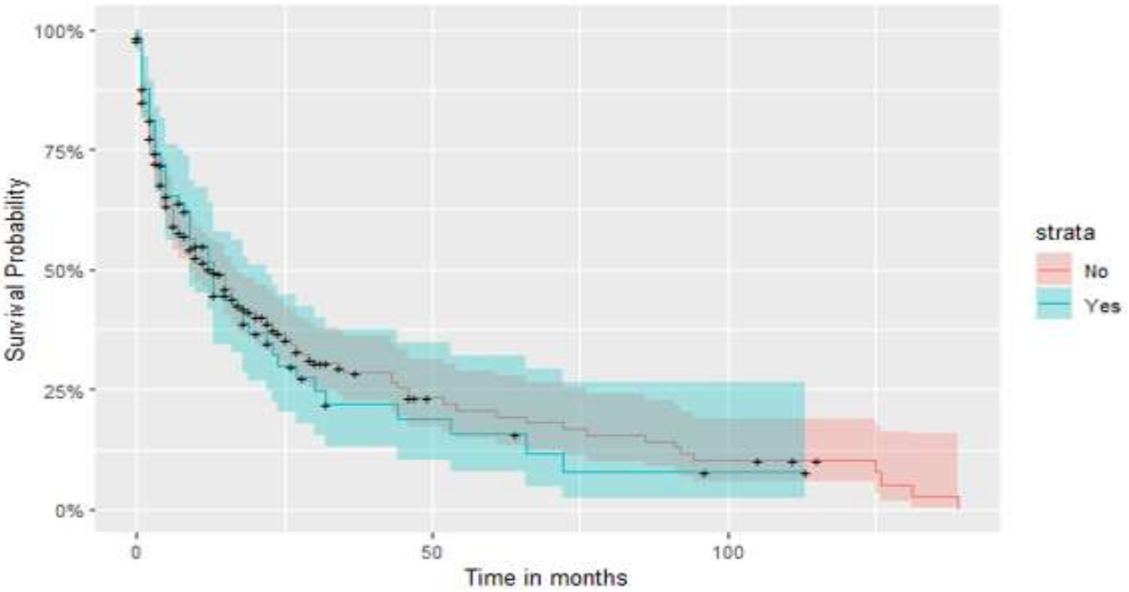


Figure 4.9: Estimate of survival curve base on alcohol consumption

Table 4.11: K-M estimate of survival time base on alcohol consumption

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Yes	89	52	58	13	1.60	9.87	15.14
No	299	182	61	14	1.94	10.21	17.80

4.2.10 Estimate of survival base on tribe

Figure 4.10 and Table 4.12 shows a distinct survival time of patients based on tribe. Hausa speaking patients were 59 of which 39 (66%) experienced the event with median survival time of 15 months. Igbos' had a total of 57 of which 30 (53%) experienced the event with 23 months median survival time. Yoruba had a total of 40 of which 23 (58%) experienced the event with median survival time of 7 months. Subjects categorised under others had a total of 232 of which 142 (61%) been event with median survival time of 13 months. However, Igbo have a 50% chance of surviving at least 23 months with less proportional deaths experienced while Yoruba patients have the least median survival time of 7 months.

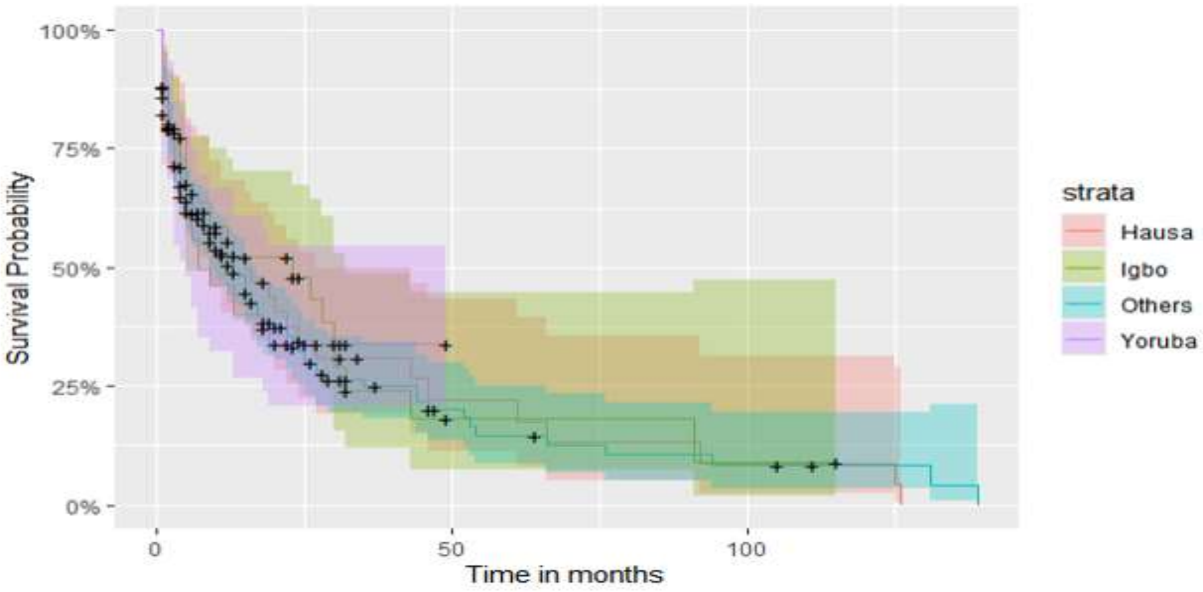


Figure 4.10: Estimate of survival curve base on tribe

Table 4.12: K-M estimate of survival time base on tribe

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Hausa	59	39	66	15	5.29	4.64	25.36
Igbo	57	30	53	23	9.37	4.63	41.37
Yoruba	40	23	58	7	3.86	0.00	14.56
Others	232	142	61	13	2.01	9.06	16.94

4.2.11 Estimate of survival base on comorbidity

Figure 4.11 and Table 4.13 shows that patients with comorbidity were 146 in total, of which 88 (60%) been event and has a median survival time of 11 months while patients with no comorbidity were a total of 242 of which 146 (60%) been events with a median survival time of 15 months. It was observed that those with comorbidity experienced less median survival time with 60 % deaths while more median survival time with 60% deaths were recorded for those with no comorbidity.

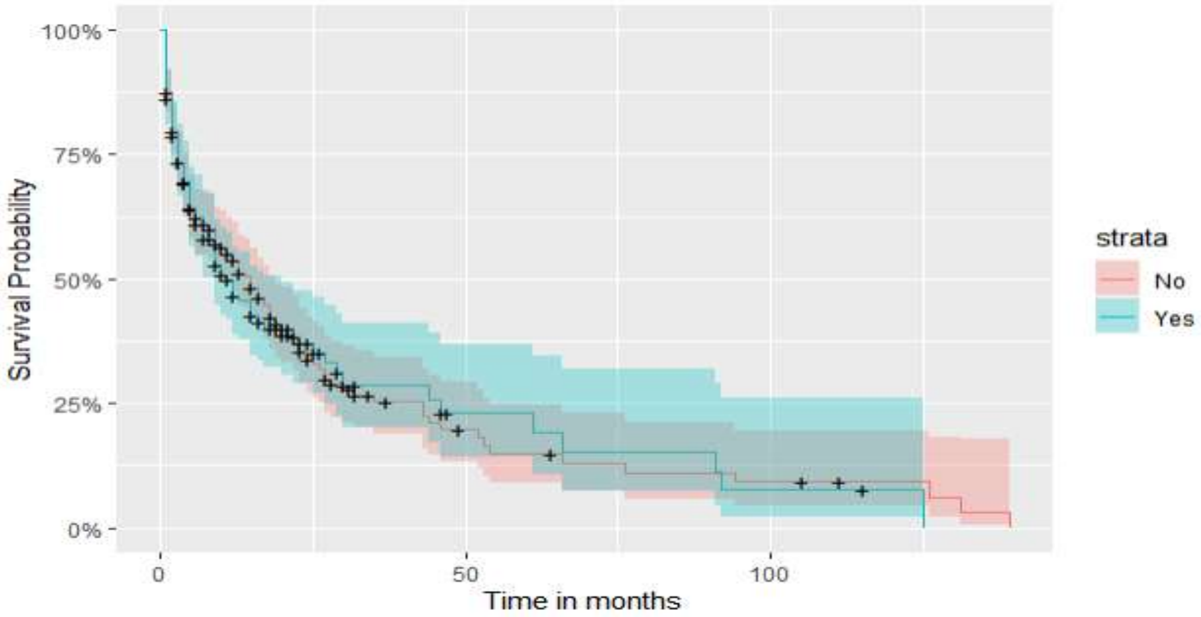


Figure 4.11: Estimate of survival curve base on comorbidity

Table 4.13: K-M estimate of survival time base on comorbidity

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Yes	146	88	60	11.15	2.32	6.45	15.55
No	242	146	60	15	2.08	10.92	19.08

4.2.12 Estimate of survival base on histology

Figure 4.12 and Table 4.14 shows a distinct survival time of patients based on histology. Forty six (46) patients had Adenocarcinoma of which 30 (65%) were event with a median survival time of 15 months. Adenosquamous have 14 patients of which 8 (57%) been events (deaths) with a median survival time of 35 months, and squamous cell carcinoma (SCC) had a total number of 328 patients of which 198 (60%) been event with a median survival time of 13. Hence, those with Adenosquamous had the highest survival time and less proportion of deaths observed while patients with Adenocarcinoma have the least survival time with more proportion of deaths observed.

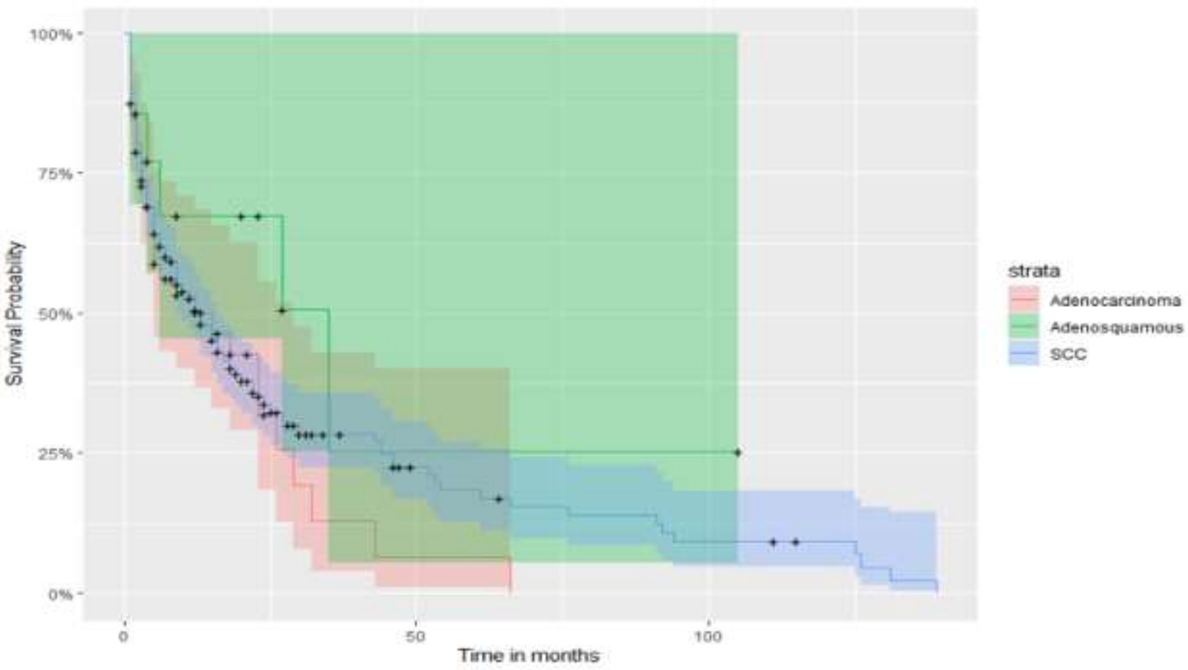


Figure 4.12: Estimate of survival curve base on histology

Table 4.14: K-M estimate of survival time base on histology

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
Adenocarcinoma	46	30	65	15	7.21	0.89	29.13
Adenosquamous	14	8	57	35	13.75	8.06	61.94
SCC	328	198	60	13	1.78	9.52	16.48

4.2.13 Estimate of survival base on tumour grade

Figure 4.13 and Table 4.15 display a distinct survival time of patients based on tumour grade. Patients with moderately differentiated (MD) tumour grade have a total number of 126 of which 74 (58%) been event with 18 months median survival time. Patients with poorly differentiated (PD) tumour grade have a median survival time of 9 months with 104 (66%) deaths observed and patients with well differentiated tumour grade (WD) have a total 102 with a median survival time of 12 months with 56 (55%) deaths recorded. Thus, patients with moderately differentiated (MD) tumour grade have highest median survival time while poorly differentiated (PD) patients have the least median survival time with more proportion of deaths 65% observed.

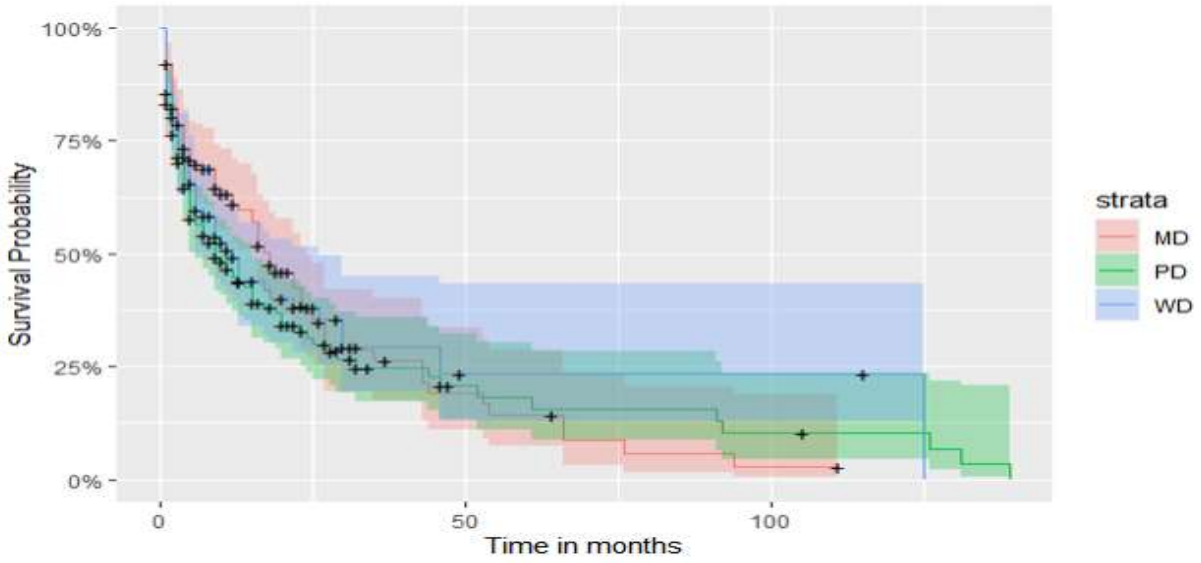


Figure 4.13: Estimate of survival curve base on tumour grade

Table 4.15: K-M estimate of survival time base on tumour grade

Category	N	Events	%Events	Median	Std. Error	0.95LCL	0.95UCL
MD	126	74	58	18	2.50	13.09	22.91
PD	160	104	65	9	2.18	4.72	13.28
WD	102	56	55	12	2.31	7.47	16.53

4.2.14 Estimate of survival base on treatment

Figure 4.14 and Table 4.16 shows that patients treated with chemo/radiotherapy have a total number of 148 of which 84 (57%) been event with a median survival time of 15 months. Those treated by chemotherapy were a total of 72 of which 48 (67%) been event with 9 months as median survival time. Radiotherapy patients had a total number of 103 of which 61 (59%) been event (deaths) with a median survival time of 16 months. Those that received neither of the treatments were a total of 65 of which 41 (63%) been event with a median survival time of 9 months recorded. However, patients treated by chemo/radiotherapy and those treated by only radiotherapy have about the same 50% survival time of 16 months with slight difference in the proportion of deaths recorded, while those treated by chemotherapy and those with no treatment have same 50% survival time of 9 months with slight difference also in the proportion of deaths observed.

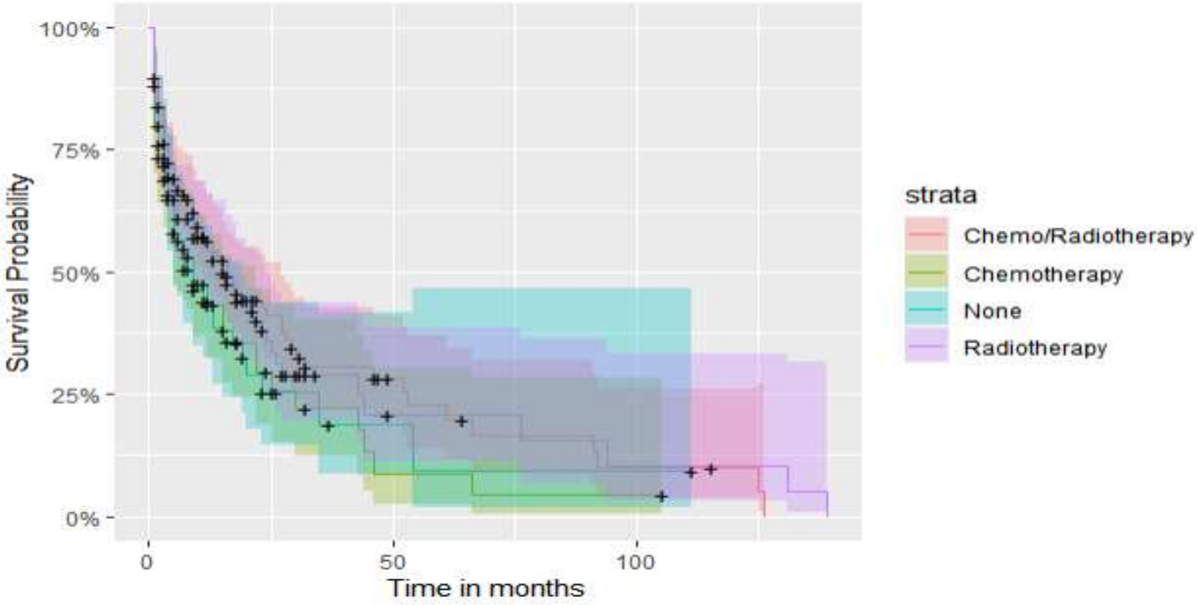


Figure 4.14: Estimate of survival curve base on treatment

Table 4.16: K-M estimate of survival time base on treatment

Category	**N**	**Events**	**%Events**	**Median**	**Std. Error**	**0.95LCL**	**0.95UCL**
Chemo/Radiotherapy	148	84	57	15	2.72	9.66	20.34
Chemotherapy	72	48	67	9	2.84	3.43	14.57
Radiotherapy	103	61	59	16	3.82	8.51	23.49
None	65	41	63	9	2.69	3.73	4.27

4.3 Log-Rank Analysis

Test for difference was performed on thirteen (13) variables used in Kaplan-Meier analysis at 5% level of significance. For each variable, H_o: No significant difference in the survival time of the categories. H_1: At least one $S_i(t)$ is different, i = 1, 2, 3, . . .

Table 4.17: Log-Rank test of difference

VARIABLES	N	OBSERVED	EXPECTED	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$	Chisq	Pval
MARITAL STATUS							
Divorced	17	9	10	0.159	0.180	1.1	0.9
Married	248	149	145	0.100	0.287		
Separated	16	9	12	0.600	0.677		
Single	10	8	7	0.167	0.185		
Widowed	97	59	60	0.015	0.023		
RELIGION							
Islam	145	93	90	0.105	0.184	0.2	0.7
Christianity	243	141	144	0.066	0.184		
FIGO							
Stage IA	2	1	3	0.964	1.04	47.4	5e-08
Stage IB	30	10	25	8.880	10.700		
Stage IIA	44	14	36	13.092	17.078		
Stage IIB	101	58	60	0.073	0.108		
Stage IIIA	63	44	34	2.690	3.399		

Table 4.17 Continued

Stage IIIB	102	81	55	12.281	17.425		
Stage IVA	33	18	18	0.003	0.003		
Stage IVB	13	8	4	4.800	5.572		
OCCUPATION							
Business	134	74	82	0.698	1.162	5.9	0.2
C/S	53	32	33	0.043	0.054		
Farmer	33	16	20	0.873	1.089		
H/W	136	96	80	3.283	5.397		
Others	32	16	19	0.549	0.646		
RECURRENCE							
Yes	35	23	37	5.32	7.19	7.2	0.007
No	353	211	197	1.00	7.19		
FAMILY HISTORY							
Yes	66	37	33	0.457	0.583	0.6	0.4
No	322	197	200.9	0.075	0.583		
SMOKING STATUS							
Yes	25	15	14	0.127	0.146	0.1	0.7
No	363	219	220	0.008	0.146		
ALCOHOL CONSUMPTION							
Yes	89	52	48	0.256	0.364	0.4	0.5
No	299	182	186	0.069	0.364		

Table 4.17 Continued

TRIBE							
Hausa	59	39	41	0.070	0.091	0.5	0.9
Igbo	57	30	33	0.244	0.305		
Yoruba	40	23	22	0.063	0.075		
Others	232	142	139	0.080	0.213		
HISTOLOGY							
Adenocarcinoma	46	30	26	0.627	0.763	2.5	0.3
Adenosquamous	14	6	10	1.736	1.940		
SCC	328	198	198	0.000	0.001		
TUMOUR GRADE							
MD	126	74	81	0.594	0.984	1.5	0.5
PD	160	104	96	0.708	1.315		
WD	102	56	57	0.030	0.042		
TREATMENT							
Chemotherapy	72	48	39	2.012	2.61	5.3	0.2
Radiotherapy	103	61	65	0.249	0.38		
Chem/Radiotherapy	148	84	95	1.356	2.48		
None	65	41	35	1.235	1.57		
COMORBIDITY							
Yes	146	88	86	0.050	0.085	0.1	0.8
No	242	146	148	0.029	0.085		

P-value of less than 0.05 was used for the significant test of Log rank test. Table 4.17 shows the p-values for FIGO (tumour stages) and recurrence to be less than 0.05, indicating statistically significant at 5%. It is therefore concluded that the survival times of patients based on the tumour stages, and recurrence of cancer are significantly different, while the survival times of patients based on marital status, religion, occupation, family history, smoking status, alcohol consumption, tribe, histology, tumour grade, treatment and comorbidity were not significantly different. However, not significantly different doesn't imply the accelerating or decelerating effect of these covariates. Hence, Accelerated Failure Time Models with four (4) distortions were fitted and the best model for the available data was selected based on AIC and BIC.

4.4 Classification and Regression Trees (CART) Model

The Classification and Regression Trees (CART) which is a machine learning algorithm was used to identify the optimal classification and predictive nodes. This form of decision tree will present the classification of patients into the event of interest (death) or censored with certain degree of sensitivity and specificity.

Table 4.18 : CART model confusion matrix

		Predicted class (training)			**Predicted class (test)**		
Actual class	**Count**	**1**	**0**	**% correct**	**1**	**0**	**% correct**
1 (Event)	234	193	41	82.5	196	38	83.8
0	154	62	92	59.7	57	97	62.9
All	388	255	133	73.5	253	135	75.5

Table 4.18 Continued

Statistics	Training (%)	Test (%)
True positive rate (sensitivity or power)	82.5	83.8
False positive rate (type I error)	40.3	37.0
False negative rate (type II error)	17.5	16.2
True negative rate (specificity)	59.7	62.9

The response variable of the cervical cancer data was classified using the classification and regression tree and split into a training data and test data to determine the predictive power of the CART model (a form of machine learning). The confusion matrix (Table 4.18) reveals correct and incorrect classified subjects as either dead or censored. In the training class, out of 234 subjects that experienced the event of interest, 41 were misclassified, resulting to 82.5% correct prediction. Similarly, out of the 154 subjects that did not experience the event of interest, 62 were misclassified leading to 59.7% correct prediction. The average percent correct classification of subjects is 73.5%. While in the test class, the average percent correct classification of subjects is 75.5%. The CART model is shown to have a power of 82.5% which is good and can be used in classifying patients with cervical cancer.

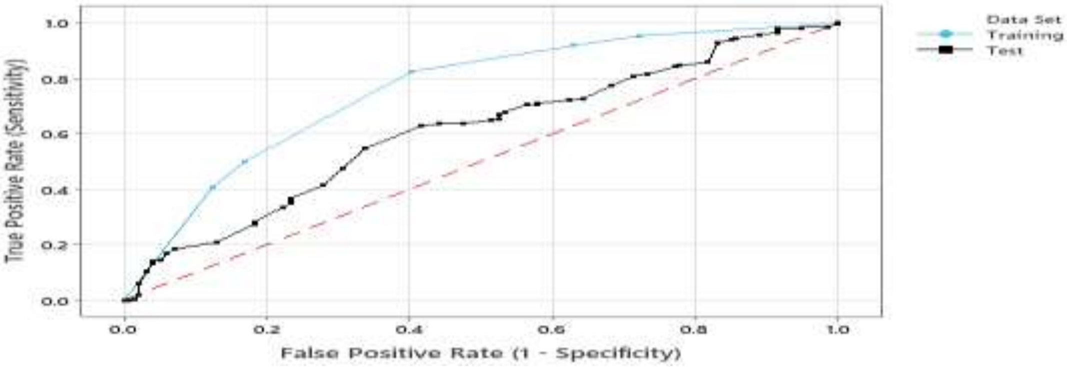


Figure 4.15: Receiver operating characteristic (ROC) curve

Table 4.19: CART model summary

Statistics	Training	Test
Average-loglikelihood	0.5604	0.8780
Area under ROC curve	0.7564	0.6091
95% CI	(0.4042, 1)	(0.5523, 0.6660)
Lift	1.3818	1.4017
Misclassification cost	0.2655	0.2448

Figure 4.15 and Table 4.19 give details of the ROC curve and loglikelihood of the training and test class. The ROC curve shows the trade-off between sensitivity and 1 − specificity. The figure which shows the AUC of 0.7564 and 0.6091 for training class and test class respectively obviously reveals that the model does well in classifying the data into patients that experience the event of interest and those censored based on the available data.

The misclassification cost versus number of terminal nodes plot was used to produce the optimal tree (Figure 4.16). The pattern keeps rising after the least misclassification cost and, thus, the tree with six terminal nodes was generated. The tree in the sequence with six terminal nodes has a misclassification cost of 0.2448.

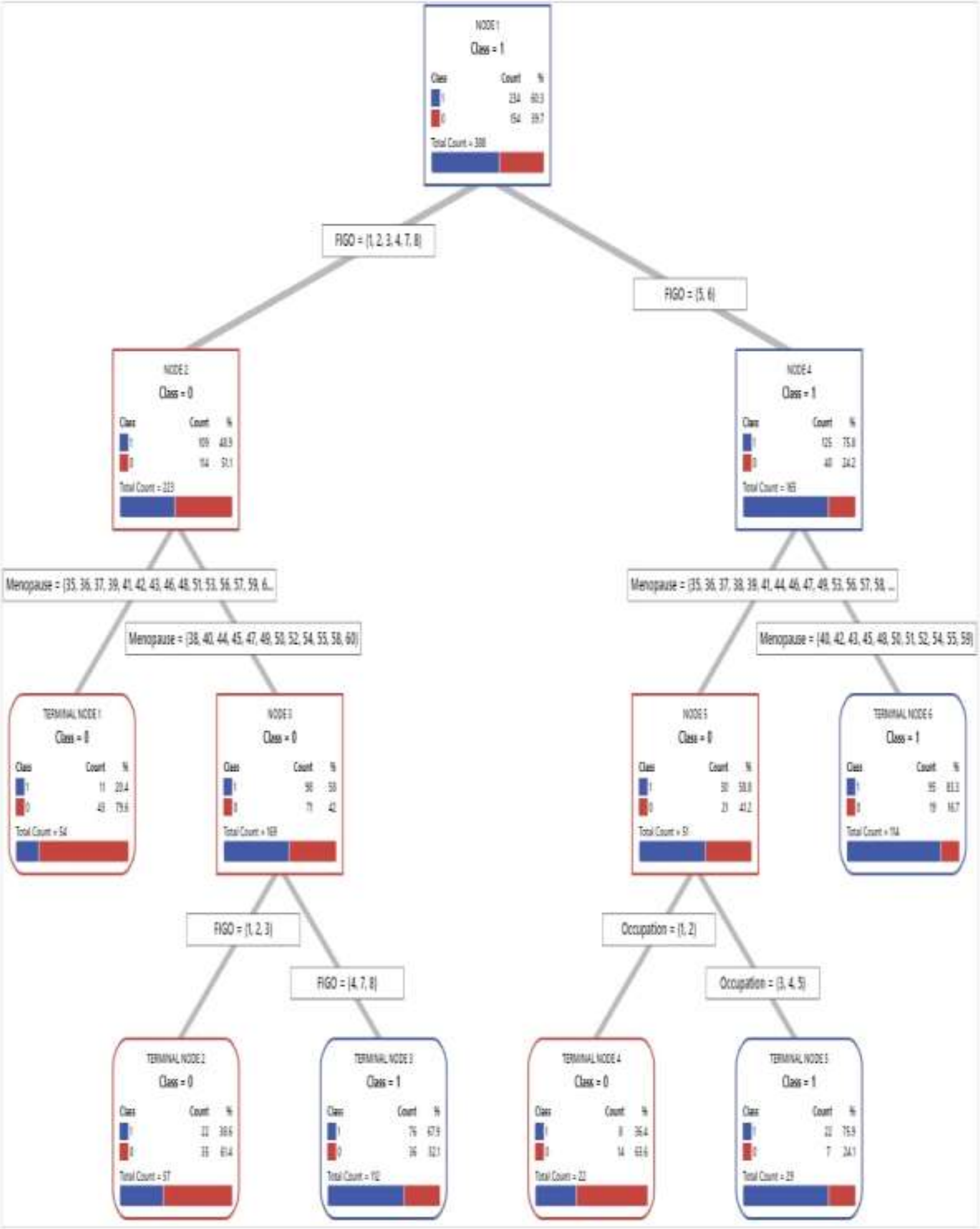


Figure 4.16: Optimal tree diagram

The classification tree (Figure 4.16) has six (6) terminal nodes. On the tree, 1, 2, 3, 4, 5, 6, 7 and 8 represent FIGO stages IA, IB, IIA, IIB, IIIA, IIIB, IVA and IVB respectively while 1, 2, 3, 4 and 5 respectively represents business, civil servants, farmers, house wife and others as occupation. Blue is for the event level (1) (death) and red is for non-event level (0) (censored). The tree diagram uses the test data set. At the root node, event (1) has 234 (60.3%) cases and non-event (0) has 154 (39.7%) cases. The root node is split using the variables FIGO stage. When FIGO stage equal 0, go to the left node (node 2) and when FIGO stage equal 1, go to the right node (node 4). Node 2 had FIGO stages IA, IB, IIA, IIB, IVA and IVB with a total of 223 cases. From the total number 223, 109 (48.9%) were death (1) and 114 (51.1%) were censored (0). Node 4 consist of FIGO stages, IIIA and IIIB with total number of 165 cases, of which 125 (75.8%) were death (1) and 40 (24.2%) were censored (0). The next splitter for left node (node 2) was menopause. Patients that attained menopause at the age of 35, 36, 37, 39, 41, 42, 43, 46, 48, 51, 53, 56, 57, 59 and 62 go to the left node (terminal node 1) and right node (node 3) had menopause ages 38, 40, 44, 45, 47, 49, 50, 52, 54, 55, 58 and 60 years.

In terminal node 1, there are total of 54 cases. From the total number of 54, 11 (20.4%) were death and (1) and 43 (79.6%) were censored (0). From the total number of 169 cases of node 3, 98 (58%) were death (1) and 71 (42%) were censored (0). Further splitter for node 3 was FIGO stages. Patients that have FIGO stages IA, IB and IIA go to the left node (terminal node 2) with a total of 57 cases of which 22 (38.6%) were death (1) and 35 (61.4%) were censored while those patients with FIGO stages IIB, IVA and IVB go to the right node (terminal node 3). Terminal node 3 had a total of 112 cases, from which 76 (67.9%) were death (1) and 36 (32.1% were censored (0). Node 4 was split into two child nodes based on menopause. Patients that reach menopause at ages 35, 36, 37, 38, 39, 41, 44, 46, 47, 49, 53, 56 and 58 go to the left node (node 5) while those that had menopause at age of 40, 42, 43, 45, 48, 50, 51, 52, 54, 55 and 59 go to the right node (terminal node 6). Node 5 had a total of 51 cases of which 30 (58.8%) were death (1) while 21 (41.2%) were censored (0). A total of 114 cases were observed in terminal node 6 out of which 95 (83.3%) were death (1) and 19 (16.7%) were censored (0).

Finally, node 5 was split into terminal node 4 and 5 based on occupation of the patients. Patients that had business and civil servant as occupation go to the left node (terminal node 4) while patients that were farmers, house wife and others go to right node (terminal node 5). In terminal node 4, there were a total of 22 cases of which 8 (36.4%) were death (1) and 14 (63.6%) were censored (0). Node 2 was the parent node to terminal node 1, node 3 to terminal node 2 and 3 and node 4 was the parent node to terminal node 6 while node 5 was to terminal nodes 4 and 5 respectively.

The ranking of terminal nodes from most pure to least pure which shows good separation of cases were node 1, 2, 4, 3, 5 and 6. Therefore, it was concluded that early menopause, FIGO stage (stage IA and IB) and occupation (business and civil servants) are the factors predictive of chance of survival for cervical cancer patients.

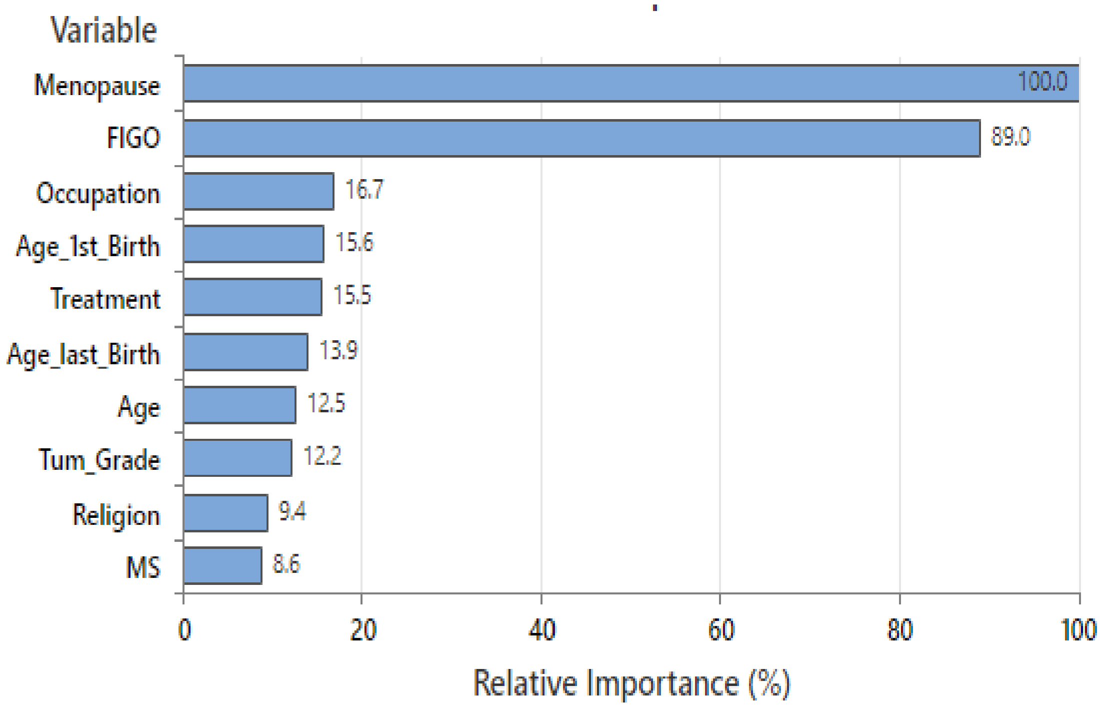


Figure 4.17: Relative variable importance

Figure 4.17 shows the relative importance of the variables. Menopause is 100% important in the prediction of cervical cancer. The FIGO stage, occupation, age at first birth, treatment exposed to and age at last birth are 89.0%, 16.7%, 15.6%, 15.5% and 13.9% respectively as important as menopause. Furthermore, age has 12.5% importance while tumour grade and religion are 12.2% and 9.4% respectively as important as menopause and the least important is the marital status with 8.6%.

Table 4.20: Optimal tree and variable importance summary

Total predictors	20
Important predictors	10
Number of terminal nodes	6
Minimum terminal node size	22

From Table 4.20, a total of twenty (20) predictors- age, parity, menopause, coitarche, marital status, FIGO stage, occupation, recurrence, age at first birth, age at last birth, smoking status, family history, menarche, alcohol consumption, tribe, histology type, tumour grade, treatment, comorbidity and religion were trained and tested but only ten (10) were relatively important variables with six optimal terminal nodes and a minimum terminal node size of twenty two (22).

4.5 Accelerated Failure Time Model Fitting

Studies by several researchers revealed that if the assumption of the Cox proportional hazard is violated, the parametric model may be adopted (Pourhoseingholi *et al*., 2007 and Moghimi-Dehkordi *et al*., 2008). Before the AFT model was fitted, the assumption of Cox proportion hazard model was tested and found to have failed, consequently, the choice of AFT model. The results of the tested assumption is given in appendix IV.

The data sets were analysed using the different AFT models such as Exponential, Weibull, Log-normal and Log-logistic models. The results from different AFT models applied to cervical cancer progression are presented in Table 4.21.

Table 4.21: Log-likelihoods, AIC and BIC in the models

Distribution	K	AIC	BIC	AIC_Wt	LL
Log-normal	39	1821.70	1844.66	0.95	-871.85
Log-logistic	39	1827.64	1850.60	0.05	-874.82
Weibull	37	1843.01	1864.79	0.00	-884.50
Exponential	38	1846.11	1868.48	0.00	-885.05

Note: K is the number of parameters, AIC_Wt is the proportion of the total predictive power, LL is the Log-likelihood of the model.

The AFT models were compared using statistical criteria - Maximum likelihood (ML) test, AIC and BIC. According to these criteria, the model with minimum AIC, BIC and higher log-likelihood value is best. The computed values of AIC and BIC for Log-normal AFT model are 1821.70 and 1844.66 respectively with 95% predictive power indicating it to be the appropriate AFT model compared to the other AFT models (Table 4.21).

Table 4.22: Estimate of prognostic factors of cervical cancer based on NHA

Variables	Coefficient value	TR	SE	Z value	P value
(Intercept)	5.462	235.57	1.537	3.55	0.00
Parity	-0.081	0.92	0.046	-1.76	0.08
Occupation (H/W)	-0.715	0.49	0.219	-3.26	0.00
Recurrence (Yes)	0.904	2.467	0.286	3.16	0.00

Table 4.22 Continued

Tribe (Yoruba)	-0.618	0.54	0.351	-1.76	0.08
Histology (ADQ)	1.005	2.78	0.507	1.98	0.05
Grade (WD)	-0.433	0.65	0.221	-1.96	0.05
Treatment (Chemo)	-0.3877	0.69	0.224	-1.68	0.09
Log (scale)	0.290	1.34	0.047	6.17	0.00

SE = Standard error

TR = Time ratio

Scale = 1.34

Log Normal Distribution

Loglikelihood (model) = -871.9

Loglikelihood (intercept only) = -924.5

Chisquare = 105.28 on 37 degrees of freedom, p = $1.8e^{-08}$

Number of Newton-Raphson iterations: 4, n = 388

Significant level considered = 0.10

The results of the lognormal AFT model fitted are presented in Table 4.22. The effect of covariate is to accelerate or decelerate the survival time of cervical cancer patients. The TR is the acceleration factor for a given covariate. A positive coefficient implies that the effect of the covariate is to prolong the survival time while a negative coefficient is to shorten the time to event (death) (Majeed, 2020). Relatively, a TR greater than 1 implies the effect of the covariate increase the survival time and otherwise decrease (“speed up”) the time to death. The TR of 0.92 for parity relative to nulliparous patients implies parity is a significant prognosis of cervical cancer and it is slightly at higher risk compared to nulliparous patients, that is, probability of a cervical cancer patient dying is 92% faster for parity. Occupation with estimate of (-0.72) = 0.49, which is a 49% increase in risk of the death for patients in house wife (H/W) category. This implies that the survival time for patients who are H/W is estimated to be 51% shorter than for patients in other categories of occupation. TR for recurrence 2.46 indicates that the survival time for cervical cancer patients who experience recurrence is longer than those without recurrence. This could be due to the fact that cervical cancer if detected early can be cured and may take longer before recurrence, that is, lengthen the survival time of the patients.

Furthermore, those with cervical cancer from Yoruba tribe have a chance of dying to be 0.54 which indicate that the survival times for Yoruba patients are estimated to be 46% lower than those patients in other categories of tribe. The estimated TR for histologic group is 2.25 indicating that longer time to death is more likely for the patients with adenosquamous cell type. This further implies that a patient diagnosed with adenosquamous will have more chance to live in the following years. The survival time for a patient diagnosed with WD tumour grade is estimated to be 65% of that of patients diagnosed with PD or MD tumour grade. Patients exposed to chemotherapy treatment survived 0.68 shorter than those exposed to radiotherapy or both. This implies that patients subjected to chemotherapy have 68% chance of dying, that is, they are at higher risk of death. Results of other insignificant variables are displayed in appendix III.

CHAPTER FIVE

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In this study, the demographic, social and clinical characterisation of three hundred and eighty eight (388) cervical cancer patients was studied. The results from the descriptive analysis showed that the data was relatively skewed and the patients' mean and median age was fifty four (54) years. Median survival time of the patients was estimated and difference between survival curves was investigated using Kaplan-Meier and Log Rank tests respectively. Findings from the Kaplan Meier showed that each patient has a 50% chance of surviving at least 13 months but not more than 17 months while log Rank test revealed a significant evidence of difference in survival times for the groups- FIGO stage and recurrence. Furthermore, the result from CART model was used to predict chance of survival of the patients. CART model indicate 82.5% power of correctly classifying patients with cervical cancer. Finally a parametric AFT model was used to identify the prognostic factors associated with cervical cancer. Exponential, Weibull, Lognormal and Log-logistic were the models considered. Based on AIC and BIC criteria, Lognormal model with least values of 1821.70 and 1844.66 respectively showed superiority to other models and was subsequently used for the analysis. Recurrence, histological type, parity, occupation, tumour grade and treatment received were found to be statistically significant at 10%.

5.2 Recommendations

Based on the findings of this research work, it is recommended that cervical cancer diagnostic and treatments facilities be made accessible and affordable to facilitate early detection. More awareness should be created on family planning to reduce multiple parity.

Human papilloma virus (HPV) vaccine should be made available in primary health care centres due to proximity of these centres to the local communities. Lognormal AFT model should be used to model the prognostic factors associated with cervical cancer. Other parametric models such as exponentiated weibull and exponentiated exponential distributions should be explored in cervical cancer and other cancer studies.

5.3 Contribution to Knowledge

The contribution to knowledge of this research work as established by the results is presented as; the research identify that life expectancy of cervical cancer patients is mostly affected by the stage (FIGO stage) of the cancer. Lognormal model performs better than Exponential, Weibull and Log-logistics models in modelling the prognostic factors for cervical cancer. Parity, occupation (house wife), tribe (Yoruba), tumour grade (Well differentiated), treatment received (Chemotherapy), recurrence and histological type were the predictive factors to deaths as a result of cervical cancer.

REFERENCES

Abd Razak, N. B. (2016). Survival Modelling, Missing Values and Frailty with Application to Cervical Cancer Data. Thesis Submitted in Partial Fulfilment of the Requirement for the Degree of Doctor of Philosophy (SCIENCE). Institute of graduate Studies, University of Malaysian Kuala Lumpur

Adamu, P. I., Adamu, M. O., Okagbue, H. I., Opoola, L. and Bishop, S. A. (2019). Survival Analysis of Cancer Patients in North Eastern Nigeria from 2004-2017-A Kaplan-Meier Method. *Macedonian Journal of Medical Sciences*, 7(4), 643-650

Ahmed, R. B., Sahar, S. M., Hamid, A. M., Mohammed, E. A., Nahid, N. & Kimiya, G. (2015).Survival Analysis of Patients with Breast Cancer Using Weibull Parametric Model. *Asian Pacific Journal of Cancer Prevention*, 16(18), 8567-8571

Akinde, O. R., Phillips, A. A., Oguntunde, O. A. & Afolayan, O. M. (2015). Cancer Mortality Pattern in Lagos University Teaching Hospital, Lagos, Nigeria. *Journal of Cancer Epidemiology,* 1-6

Akinremi, T. O., Nazeer, S. & Totsch, M. (2005). Reduced Alcohol use in the Staining of Papsmears: A Satisfactory Low-Cost Protocol for Cervical Cancer Screening. *Acta Cytology*, 49(2), 169-172

Alfensi, F. (2018). The Comparison of Proportional Hazards and Accelerated Failure Time Models in Analysing the First Birth Interval Survival Data. *Journal of Physics*, 1-10

American Cancer Society (2014). Cancer Facts and Figures 2014. American Cancer Society (ACS) Atlanta GA:

Anfinan, N. & Sait, K. (2020). Indicators of Survival and Prognostic Factors in Women Treated with Cervical Cancer at a Tertiary Care Center in Saudi Arabia. *Annual Saudi Medical*, 40(1), 25-35

Anorlu, R. I., Obodo, K. & Makwe, C. C. (2010). Cancer Mortality Among Patients Admitted to Gynaecological Wards t Lagos University Teaching Hospital, Nigeria. *International Journal of Gynaecology and Obstetrics*, 110(3), 268-269

Anto, J. V. & George, L. (2019). Survival Modelling and Analysis for Time to Failures of Aircraft Glass. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, 6(2), 252-257

Anya, S. E., Ezugwu, F. O. & Okaro, J. M. (2006). Gynaecologic Mortality in Enugu, Nigeria. *Tropical Doctors*, 36(4), 235-236

Anyasi, H. I. & Foss, A. M. (2021). A Comparative Analysis of Cervical Cancer Prevention between Nigeria and Nordic Countries that have Experience a Decline in Cervical Cancer Incidence. *International Health*, 13, 307-317

Atlam, M., Torkey, H., El-Fishawy, N. & Salem, H. (2021). Coronavirus Disease 2019 (COVID-19): Survival Analysis Using Deep Learning and Cox Regression Model. *Pattern Analysis and Applications*, 24, 993-1005

Awodele, O., Adeyomoye, A. A. A., Awodele, D. F., Kwashi, V., Awodele, I. O. & Dolapo, D. C. (2011). A Study on Cervical Cancer Screening Amongst Nurses in Lagos University Teaching Hospital, Lagos, Nigeria. *Journal of Cancer Education*, 26 (3), 497 - 504.

Ayaneh, M. G., Dessie, A. A. & Ayele, A. W. (2020). Survival models for the Analysis of Waiting Time to First Employment of New Graduates: A Case of 2018 Debre Markos University Graduates, North West Ethiopia. *Education Research International*, 1-10

Bewick, V., Cheek, L. & Ball, J. (2004). Statistics Review 12: Survival Analysis. Critical Care, London, 8, 389-394

Bolarinwa, I. A. & Micheal, V. A. (2020). Parametric Survival Modelling of Tuberculosis Data - A Case Study of Federal Medical Centre Bida, Nigeria. *Modern Applied Science*, 14(7), 37-49

Bosch, F. X., Manos, M. M., Muñoz, N., Sherman, M., Jansen, A. M. & Peto, J. (1999). Prevalence of Human Papilloma-Virus in Cervical Cancer: A Worldwide Perspective. *Journal National Cancer Institute*, 87(11), 796-802

Bradburn, M., Clark, T., Love, S. & Altman, D. (2003). Survival Analysis Part II: Multivariate Data Analysis – An Introduction to Concepts and Methods. *British Journal of Cancer*, 89, 431-436

Brandt, V., Sioulas, D., Basaran, D., Kuhn, T., La Vigne, K., Gardner, G. J., Sonada, Y., Chi, D. S., Long Roche, K. C., Mueller, J. J., Jewell, E. L., Broach, V. A., Zivanovic, O., Abu-Rustum, N, R. & Leitao Jr M. M. (2019). Minimally Invasive Surgery versus Laparotomy for Radical Hysterectomy in the Management of Early-Stage Cervical Cancer: Survival Outcomes. *Gynecologic Oncology*, 1-7

Bray, F. (2014). Transitions in Human Bevelopment and the Global Cancer Burden. In: Wild, C.P., Stewart, B. (Eds.), World Cancer Report 2014. International Agency for Research on Cancer, Lyon.

Cao, N., Zhao, A., Zhao, G., Wang, X., Han, B., Lin, R., Zhao, Y. & Yang, J. (2015). Survival Analysis of 272 Patients with Pancreatic Cancer Undergoing Combined treatment. *Integrated Cancer Therapies*, 14(2), 133-139

Carter, B. B., Zhang, Y., Zou, H., Zhang, C., Zhang, X., Sheng, R., Qi, Y., Kou, C. & Li, Y. (2021). Survival analysis of Patients with Tuberculosis and Risk Factors for Multidrug-Resistant Tuberculosis in Monrovia, Liberia. *Plos One*, 16(4), 5-8

Chakraborty, S. (2018). A Step-wise Guide to Performing Survival Analysis. *Cancer Research Statistics Treat*, 1, 41-45.

Cheah, P. L.& Looi, L. M. (1999). Carcinoma of the Uterine Cervix. A review of its pathology and Commentary on the Problem in Malaysians. *Malaysian Journal of pathology*, 21(1), 1-15

Chen, H.-H., Meng, W.-Y., Li, R.-Z., Wang, Q.-Y., Wang, Y.-W., Pan, H.-D., Yan, P.-Y., Wu, Q-B., Liu, L., Yao, X.-J., Kang, M. & Leung, E. L.-H. (2021). Potential Prognostic Factors in Progression Free Survival for Patients with Cervical Cancer. *BMC Cancer*, 21, 531-541

Clark, T. G., Bradburn, M. J., Love, S. B. & Ahman, D. G. (2003). Survival Analysis Part I: Basic Concepts and First Analyses. *British Journal of Cancer*, 89, 232-238

Daraba, G., Brigitte, S. C. & Jaba, E. (2017). Estimation of Unemployment Duration in Botosani County Using Survival Analysis. *Economics Science Series*, 1, 155-161

Denny, L. (2011). Cervical Cancer Treatment in Africa. *Current Opinion Oncology,* 23, 469-474.

De Oliveira, C. Watt, R. & Hamer, M. (2010). Tooth Brushing, Inflammation and Risk of Cardiovascular Disease: Results from Scottish Health Survey. *British Medical Journal*, 340, 24-31

De Sanjose, S., Quint, W. G, Alemany, L., Geraets, D. T., Klaustermeier, J. E., Lloveras, B., Tous, S., Felix, A., Bravo, L. E., Shin, H.-R. & Vallejos, C. S. (2010) Human Papillomavirus Genotype Attribution in Invasive Cervical Cancer: A Retrospective Cross-sectional Worldwide Study. *Lancet Oncology*, 11, 1048-1056

Dumville, J. C., Worthy, G., Bland, J. M. Cullum, N., Dowson, C., Iglesias, C., Mitchel, J. L., Nelson, E. A., Soares, M. O. & Torgerson, D. J. (2009). Larval Therapy for Leg Ulcers (Venus II): Randomised Controlled Trial. *British Medical Journal*, 338, 773-779

Elisa, T. L. & John, W. W. (2003). Statistical Methods for Survival Data Analysis. A JOHN WILEY & SONS, INC., PUBLICATION

Elmajjaoui, S., Ismaili, N., El Kacemi, H., Kebdani, T., Sifat, H., & Benjaafar, N. (2016). Epidemiology and Outcome of Cervical Cancer in National Institute of Morocco. *BMC Womens Health*, 16, 1-8.

Emmert-streib, F. & Dehmer, M. (2019). Introduction to Survival Analysis in Practice. *Machine Learning and knowledge Extraction*, 1, 1013-1038

Eryurt, M. A. & Koc, I. (2012). Internal Migration and Fertility in Turkey: Kaplan-Meier Survival Analysis. *International Journal of Population Research*, 20-22

Etikan, I., Abubakar, S. & Al-Kassim, R. (2017). The Kaplan-Meier Estimate in Survival Analysis. *Biometric and Biostatistics International Journal*, 5(2), 10.15406/bbij.2017.05.00128

Fagbamigbe, A. F. & Idemuda, E. S. (2016). Survival Analysis and Prognostic Factors of Timing of First Child Birth among Women Nigeria. *BioMed Central Pregnancy and Child Birth*, 16(102), 1-12

Fadnavis, R., A. (2019). Application of Machine Learning for Survival Analysis- A Review. IOSR J Eng (IOSRJEN), 9(5), 56-60

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C. & Parkin, D. M. (2010). Estimates of Worldwide Burden of Cancer in 2008: GLOBACAN 2008. *International Journal of Cancer*, 127(12), 2893-2917

Flynn, R. (2012). Survival Analysis. *Journal of Clinical Nursing*, 1-9

Franco, E. L., Rohan, T. E. & Villa, L. L. (1999). Epidemiologic Evidence and Human Papillomavirus Infection as a Necessary Cause of Cervical Cancer. *Journal National Cancer Institute*, 91(6), 506-511

Frempong, N. K., Osei-Mensah, C., Asamoah, O. D. & Okyere, E. (2012). Survival Analysis on Marriage and Divorce in the Kumasi Metropolis. *Canadian Journal on Computing in Mathematics, Natural Sciences, Engineering and Medicine*, 3(5), 159-163

Ghodrati, B. & Uday, K. (2005). Reliability and Operating Environment-based Spare Parts Estimation Approach: A Case Study in Kiruna Mine. *Sweden Journal of Quality Maintenance Engineering*, 11(2), 169-184

Gizaw, M., Addissie, A., Getachew, S., Ayele, W., Mitiku, I., Moelle, U., Yusuf, T., Begoihn, M., Assefa, M., Jemal, A. & Kantelhardt, E. J. (2017). Cervical Cancer Patients Presentation and Survival in the only Oncology Referral Hospital, Ethiopia: A Retrospective Cohort Study. *Infect Agent Cancer*, 12, 1-7

Gogtay, N. J. & Thatte, U. M. (2017). Survival Analysis. *Journal of the Association of Physicians of India*, 65, 80-84

Goldie, S. J., Grima, D., Kohli, M., Wright, T. C., Weinstein, M. & Franco, E. (2003). A Comprehensive Natural History Model of HPV Infection and Cervical Cancer to Estimate the Clinical Impact of a Prophylactic HPV-16/18 Vaccine. *International Journal of Cancer*, 106(6), 896-904.

Gurmu, S. E. (2018). Assessing Survival Time of Women with Cervical Cancer using Various Parametric Frailty Models: A Case Study at Tikur Anbessa Specialised Hospital Addis Ababa, Ethiopia. *Ann Data*, 9-13

Harrison, M. L., Gore, M. E., Spriggs, D., Kaye, S., Lasonos, A., Hensley, M., Aghajanian, C., Venkatraman, E. & Sabbatini, P. (2007). Duration of Second or Greater Clinical Remission in Ovarian Cancer: Exploring Potential Endpoints for Clinical Trials. *Gynecologic Oncology*, 106(3), 469-475

Haughton, D. & Haughton, J. (2011). Living Standards Analysis. Newyork: Statistics for Social and Behavioural Sciences

Ho, C. M., Chien, T. Y., Huang, S. H., Wu, C. J., Shih, B. Y. & Cheng, S. C. (2004). Multivariate Analysis of Prognostic Factors and Outcomes in Early Cervical Cancer Patients Undergoing Radical Hysterectomy. *Gynaecologic Oncology*, 93, 458-464

IARC (2007). *Human Papillomavirus Monogr Eval Carcinog Risks Hum.* 90, 1-636. PMID: 18354839. Retrieve from http://publications.iarc.fr/108

Ilevbare, O. E., Adegoke, A. A. & Adelowo, C. M. (2020). Drivers of Cervical Cancer Screening Uptake in Ibadan, Nigeria. *Heliyon*, 6, 1-5

Jedy-Agba, E., Curado, M. P., Ogunbiyi, O., Oga, E., Fabowale, T., Igbinoba, F., Osubor, G., Otu, T., Kumai, H., Koechlin, A., Osinubi, P., Dakum, P., Blattner, W. & Adebamowo, C. A. (2012). Cancer Incidence in Nigeria: A Report from Population-based Cancer Registries. *Cancer Epidemiology*, 36, 271-278

Jensen, P. T. (2007). Gynaecological Cancer and Sexual Functioning: Does Treatment Modality Have an Impact? *Sexologies*, 16(4), 279-285

Johnson, L. L. (2018). An Introduction to Survival Analysis. *Principles and Practice of Clinical Research*, 373-381

Kantelhardt, E. J., Moelle, U., Begoihn, M., Addissie, A., Trocchi, P., Yonas, B., Hezkiel, P., Stang, A., Thomssen, C., Vordermark, D., Gemechu, T., & Gebrehiwot, Y. (2014). Cervical Cancer in Ethiopia: Survival of 1,059 Patients Who Received Oncologic Therapy. *Oncologist*, 19, 727-734

Kartsonaki, C. (2016). Survival Analysis. *Medical Statistics*, 263-270

Khaemba, N. E., Mugo, C. W. & Mutai, C. (2013). The Survival of Patients with Cancer of the Cervix in Nairobi, Kenya. *African Journal of Health Sciences*, 25(2), 92-103

Kidanto, H. L., Kilewo, C. M. & Moshiro, C. (2002). Cancer of the Cervix: Knowledge and Attitudes of Female Patients Admitted at Mwuhimbili National Hospital, Dar es Salaam. *East Africa Medical Journal*, 79, 467 - 469.

Kitabo, C. A. & Kim, J. T. (2014). Survival Analysis of Loan Repayment Rate for Customers of Hawassa Commercial Bank of Ethiopia. *Journal of the Korean Data and Information Science Society*, 25(6), 1591-1598

Kleinbaum, D. G. & Klein, M. (2012). Survival Analysis: A Self-Learning. In Gail, M., Krickeberg, K., Samet, J. M., Tsiatis, A. & Wong, W. (Eds.), *Statistics for Biology and Health* (pp. 8-11). London: Springer Science

Ma, Z. S. & Krings, A. W. (2008). Survival Analysis Approach to Reliability, Survivability and Prognostic and Health Management (PHM). Proceedings of IEEE-AIAA Aerospace Conference. Bigsky, Montana, 101, 1-8

Ma, Z. S. (2021). A Unified Survival Analysis Approach to Insect Population Development and Survival Times. *Scientific Reports*, 11, 8223-8234

Majeed, A.- F. (2020). Accelerated Failure Time Models: An Application in Insurance Attrition. *The Journal of Risk Management and Insurance*, 1-18

Marshall, R. J. (2001). The use of Classification and Regression Trees in Clinical Epidemiology. *Journal of Clinical Epidemiology*, 54, 603-609

Mascarello, K. C., Zandonade, E. & Amorim, M. H. C. (2013). Survival Analysis of Women with Cervical cancer Treated at a Referral Hospital for Oncology in Espirito Santo State, 2000-2005. *Cad Saúde Pública, Rio de Janeiro*, 29(4), 823-831

Mavri, M. & Ioannou, G. (2008). Customer Switching Behaviour in Greek Banking Services Using Survival Analysis. *Managerial Finance*, 34(3), 186-197

Mills M. (2011). Introducing Survival and Event History Analysis. Illustration ed. London: SAGE Publication.

Moghimi-Dehkordi, B., Safaee, A., Pourhoseingholi, M. A., Fatemi, R., Tabeie, Z. & Zali, M. R. (2008). Statistical Comparison of Survival Models for Analysis of Cancer Data. *Asian Pacific Journal of Cancer Prevention*, 9, 417-420.

Mohammedi, A., Jalili-Ghazizadeh, M., Moslehi, I. & Yousefi-Khoshqalb, E. (2020). Survival analysis of Water Distribution Network under Intermittent Water Supply Conditions. *Water Supply*, 1-11.

Moore Higgs, G. J. & Chafe, S. M. (2001). Outcomes in Radiation Therapy Multidisciplinary Management. Sudbury: Jones and Barllett Publishers, Inc.

Musa, J., Nankat, J., Achenbach, C. J., Shambe, I. H., Taiwo, B. O., Mandong, B., Daru, P. H., Murphy, R. L. & Sagay, A. S. (2016). Cervical Cancer Survival in a Resource Limited setting- North Central Nigeria. *Infectious Agents and Cancer*, 11(15), 1-7

Mwaka, A. D., Garimoi, C. O., Were, E. M., Roland, M., Wabinga, H. & Lyratzopoulos, G. (2016). Social, Demographic and Healthcare Factors Associated with Stage at Diagnosis of Cervical Cancer: Cross-sectional Study in a Tertiary Hospital in Northern Uganda. *BMJ Open*, 6, 1-9

National Cancer Institute (US). Cancer Stat Facts: Cervical Cancer [Internet]. Bethesda, MD: National Cancer Institute; 2018

Okoye, C. A. (2014). Histopathological Pattern of Cervical Cancer in Benin City, Nigeria. *Journal of Medical Investment Practice*, 9 (4), 147 - 150.

Okwor, V. C., Fagbamigbe, A. F. & Fawole, O. I. (2017). Survivorship of Patients with Head and Neck Cancer Receiving Care in a Tertiary Health Facility in Nigeria. *Cancer Management and Research*, 9, 331-338

Ozturk, S., Fthenakis, V. & Faulstich, S. (2018). Assessing the Factors Impacting on Reliability of Wind Turbines via Survival Analysis. Energies, 11, 3034-3053

Parkin, D. M. (2006). The Evolution of the Population-based Cancer Registry. *National Review Cancer*; 6, 603-612

Perera, M. & Dwivedi, A. K. (2019). Statistical Issues and Methods in Designing and Analysing Survival Studies. *Cancer Report*, 1176

Piñeros, M., Znaor, A., Mery, L. & Bray, F. (2017). A Global Cancer Surveillance Framework within Non Communicable Disease Surveillance: Making the Case for Population-Based Cancer Registries. *Epidemiology Review*, 39, 161-169

Ponnuraja, C. & Venkatesan, P. (2010). Survival Models for Exploring Tuberculosis Clinical Trial Data- An Empirical Comparison. *Indian Journal of Science and Technology*, 3(7), 755-758

Pourhoseingholi, M. A., Hadizadeh, E., Morghimi-Dehkordi, B., Safaee, A., Abadi, A. & Zali, M. R. (2007). Comparing Cox Regreession and Parametric Models for Survival of Patients with Gastric Carcinoma. *Asian Pacific Journal of Cancer Prevention*, 8, 412-416

Prendiville, W. & Sankaranarayanan, R. (2017). Colposcopy and Treatment of Cervical Precancer. International Agency for Research and Cancer Technical Publication No 45

Prinja, S., Gupta, N. & Verma, R. (2010). Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine*, 35(2), 217-221

Radstone, D. & Kunkler I. H. (2003). Cervix, Body of Uterus, Ovary, Vagina, Vulva, Gestational Trophoblast Tumours. In Bomford, C. K., Kunkler, I. H., Miller, H. and Walter, J. (Eds.), *Walter and Miller's Book of Radiotherapy Radiation Physics, Therapy and Oncology* (PP. 465-486)

Ren, F., Zhang, J., Gao, Z., Zhu, H., Chen, X., Liu, W., Xue, Z., Gao, W., WU, R., Yi, L. V. & Hu, L. (2018). Racial Disparities in the Survival Time of Patients with Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma between Chinese Patients and Patients of other Racial Groups: A Population-Based Study from 2004 to 2013. *Oncology Letters*, 16, 7102-7116

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. J. C., Nussenbaum, B. & Wang, E. W. (2010).a Practical Guide to Understanding Kaplan-Meier Curves. *Otolaringo Head, Neck and Surgery*, 143(3), 331-336

Sabbatini, P. & Spriggs, D. R. (2006). Consolidation for Ovarian Cancer in Remission. *Journal of Clinical Oncology*, 24(4), 537-539

Saikia R. & Barman, M. P. (2017). Comparing Accelerated Failure Time Models with its Specific Distributions in the Analysis of Esophagus Cancer patients Data. *International Journal of Computational and Applied Mathematics*, 12(2), 411-424

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martinez-Valverde, S., Toledano-Toledano, F. & Garduña-Espinosa, J. (2020). A Survival Analysis of Covid-19 in the Mexican Population. *BMC Public Health*, 20, 1-8

Sanizah, A., Hasfariza, F., Norin Rahayu S. & Nur Niswah Naslina, A. (2014). Determinants of Marital Dissolution: A Survival Analysis Approach. *International Journal of Economics and Statistics*, 2, 348-354

Sannachi, L., Gangeh, M., Tadayyon, H., Sadeghi-Naini, A., Gandhi, S., Wright, F. C., Slodkowska, E., Curpen, B., Tran, W. & Czarnota, G. J. (2018). Response Monitoring of Breast Cancer Patients Receiving Neoadjuvant Chemotherapy using Quantitative Ultrasound, Texture and Molecular Features. *Journal of Oncology*, 1-4

Saroj, R. K., Murty, K. N. & Kumar, M. (2018). Survival Analysis for under Five Child Mortality in Uttar Pradesh. *International Journal of Research and Analytical Reviews*, 5(3), 782-786

Schiffman, M., Brinton, L. A., Devesa, S. S. & Fraumeni Jr., J. F. (1996). Cervical Cancer. In Schottenfeld D., Fraumeni J. F., Jr (Eds.), *Cancer Epidemiology and Prevention. New York* (PP. 109-116), USA: Oxford University Press

Seungyeoun, L. & Heeju, L. (2019). Review of Statistical Methods for Survival Analysis Using Genomic Data. *Genomic and Informtics*, 17(4), 41-43

Severe, P., Juste, M. A., Ambroise, A., Eliacin, L., Marchand, C., Apollon, S., Edward, A., Bang, H., Nicotera, J., Godfrey, C., Gulick, R. M., Johnson, W. D., Pape, J. W. & Fitzgerald, D. W. (2010). Early Versus Standard Antiretroviral Therapy for HIV-Infected Adults in Haiti. *New England Journal of Medicine*, 363, 257-265

Stevenson, M. (2007). An Introduction to Survival Analysis. *EpiCentre*, 1-31

Sun, Y., Gong, J., Guo, B., Shang, J., Cheng, Y. & Xu, H. (2018). Association of Adjuvant Radioactive Iodine Therapy with Survival in Node-Positive Papillary Thyroid Cancer. *Oral Oncology*, 87, 152-157

Takahashi, Y., Matsomoto, A., Morisaki, K. & Omura, S. (2006). Patulibacter Minatonensis Gen. Nov., Sp. Nov. A Novel Actinobacterium Isolated using an Agar Medium Supplemented with Superoxide Dismutase, and Proposal of Patulibacteraceae Fam. Nov. *International Journal of Systematic and Evolutionary Microbiology*, 56, 401-406

Tesfay, B., Getinet, T. & Derso, E. A (2021). Survival Analysis of Time to Death of Breast Cancer Patients: In Case of Ayder Comprehensive Specialized Hospital Tigray, Ethiopia. *Cogent Medicine*, 8(1), 7-12

Thuijs, D. J. F. M., Hickey G. L. & Osnabrugge R. L. J. (2018). Statistical Primer: Basics of Survival Analysis for the Cardiothoracic Surgeon. *Interactive CardioVascular and Thoracic Surgery*, 27, 1-4

Tittonell, P., Vanlauwe, B., Corbeels, M. & Giller, K. E. (2008). Yield Gaps Nutrient use Efficiencies and Response to Fertilizers by Maize across Heterogeneous Smallholder Farms of Western Kenya. *Plant Soil*, 313, 19-37

Tolley, H. D., Barnes, J. M. & Freeman, M. D. (2016). Survival Analysis. *Forensic Epidemiology*, 262-284

Torre, L. A., Bray, F. Siegel, R. L., Ferlay, J., Lortet-Tieulent, J. & Jemal A. (2015). Global Cancer Statistics, 2012. *A Cancer Journal for Clinicians*, 65(2), 87-108

Versmissen, J., Oosterveer, D. M., Yazdanparah, M., Defesche, J. C., Basart, D. C., Liem, A. H., Heeringa, J., Witteman, J. C., Lansberg, P. J., Kastellen, J. J. & Sijbrands, E. J. (2008). Efficacy of Strain in Familial Hypercholesterolaemia: A Long term cohort Study. *British Medical Journal*, 337, 24-33

Vishma, B. K., Prakash, B., Praveen, K. & Renuka M. (2016). Survival and Prognostic Factors for Cervical Cancer: A Hospital Based Study in Mysuru India. *International Journal of Community Medicine and Public Health*, 3(1), 218-223

Waggoner, S. E. (2003). Cervical cancer. *The Lancet*, 361(9376), 2217-2225

Walboomers, J. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A. & Shah K. V. (1999). Human Papillomavirus is a Necessary Cause of Invasive Cancer Worldwide. *Journal of Pathology*, 189(1), 12-21

Wassie, M., Argaw, Z., Tsige, Y., Abebe, M. & Kisa, S. (2019). Survival Status and Associated Factors of Death among Cervical Cancer Patients Attending the Tikur Anbesa Specialisd Hospital, Addis Ababa, Ethiopia: A Retrospective Cohort Study. *Biomedcentral*, 19, 1221-1231

WHO, (2019). Global strategy towards the elimination of cervical cancer as a public health problem.https://www.who.int/docs/defaultsource/documents/cervical-cancer-elimination-draft-strategy.pdf

WHO, (2018). WHO Director-General calls for all countries to take action to help end the suffering caused by cervical cancer. https://www.who.int/reproductivehealth/call-to-action-eliminationcervical-cancer/en/

Yoosefi, M., Ahmad, R. B., Naghmeh, K., mohammsd, A., Pourhoseingholi, Alireza, A. B. & Azink (2018). Survival Analysis of Colorectal Cancer Patients Using Exponentiated Weibull Distribution. *International Journal of Cancer Management*, 11(3), 86-89

Yu, T.-J., Liu, Y.-Y., Hu, X. & Di, G.-H. (2018). No Survival Improvement of Contralateral Prophylactic Mastectomy among Women with Invasive Lobular Carcinoma. *Journal of Surgical Oncology*, 1-8

Yu, L., Gong, H., Li, Q., Ren, H., Wang, Y., He, H., Li, T., & Song, Q. (2021). Survival Analysis of Radiation Therapy in Ovarian Cancer: A SEER Database Analysis. *Journal of Oncology*, 1-11

APPENDIX I

Misclassification

Input Misclassification Cost	predicted Class	
Actual Class	1	0
1		1.00
0	1.00	

		Training			Test		
Actual Class	**Count**	**Misclassed**	**%Error**	**Cost**	**Misclassed**	**%Error**	**Cost**
1 (Event)	234	41	17.5	0.1752	85	36,3	0.3632
0	154	62	40.3	0.4026	68	44,2	0.4416
All	388	103	26.5	0.2889	153	39.4	0.4024

APPENDIX II

Model selection based on AICc:

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
Lognormal	39	1821.70	0.00	0.95	0.95	-871.85
Loglogistic	39	1827.64	5.94	0.05	1.00	-874.82
Weibull	37	1843.01	20.38	0.00	1.00	-884.50
Exponential	38	1846.11	23.93	0.00	1.00	-885.05

APPENDIX III

Output of insignificant variables from Log normal analysis

	Value	**Std. Error**	**z**	**p**
(Intercept)	5.46165	1.53734	3.55	0.00038
Age	0.00830	0.00886	0.94	0.34883
msMarried	-0.16774	0.41100	-0.41	0.68318
msSeperated	0.40319	0.57063	0.71	0.47984
msSingle	-0.46934	0.62315	-0.75	0.45135
msWidowed	-0.13759	0.44180	-0.31	0.75548
relIslam	0.19086	0.21011	0.91	0.36366
Parity	-0.08112	0.04642	-1.75	0.08057
figoStage I B	0.60454	1.28811	0.47	0.63884
figoStage II A	0.46666	1.27199	0.37	0.71371
figoStage II B	-0.77923	1.24180	-0.63	0.53033
figoStage III A	-1.33031	1.24300	-1.07	0.28451
figoStage III B	-1.34937	1.24842	-1.08	0.27976
figoStage IV A	-1.12830	1.27001	-0.89	0.37432
figoStage IV B	-2.11859	1.30948	-1.62	0.10569
occC/S	-0.30232	0.26132	-1.16	0.24732
occFarmer	0.06390	0.32324	0.20	0.84328
occH/W	-0.71544	0.21887	-3.27	0.00108
occOthers	0.35083	0.32432	1.08	0.27937
Age_1st_Birth	-0.03047	0.02881	-1.06	0.29013
Age_last_Birth	0.00125	0.01549	0.08	0.93561
recYes	0.90442	0.28575	3.17	0.00155
famhistYes	-0.11322	0.21833	-0.52	0.60406
smokstYes	-0.03614	0.36993	-0.10	0.92218
alcYes	-0.07775	0.21986	-0.35	0.72359
Menarche	-0.02362	0.05384	-0.44	0.66083
Coitarche	-0.02709	0.03797	-0.71	0.47553
tribeIgbo	-0.44584	0.35721	-1.25	0.21199

```
tribeOthers          -0.37548       0.26598          -1.41             0.15805
tribeYoruba          -0.61801       0.35075          -1.76             0.07808
histAdenosquamous     1.00528       0.50679           1.98             0.04730
histSCC               0.18599       0.24410           0.76             0.44611
gradePD              -0.12692       0.19629          -0.65             0.51788
gradeWD              -0.43308       0.2210           -1.96             0.05005
trtChemotherapy      -0.37718       0.22357          -1.69             0.09158
trtNone              -0.21828       0.23888          -0.91             0.36083
trtRadiotherapy      -0.07742       0.20485          -0.38             0.70548
comorYes             -0.20343       0.16788          -1.21             0.22560
Log(scale)            0.29028       0.04713           6.16             7.3e-10
Scale= 1.34

Log Normal distribution

Loglik(model)= -871.9   Loglik(intercept only)= -924.5

Chisq= 105.28 on 37 degrees of freedom, p= 1.8e-08

Number of Newton-Raphson Iterations: 4

n= 388
```

APPENDIX IV

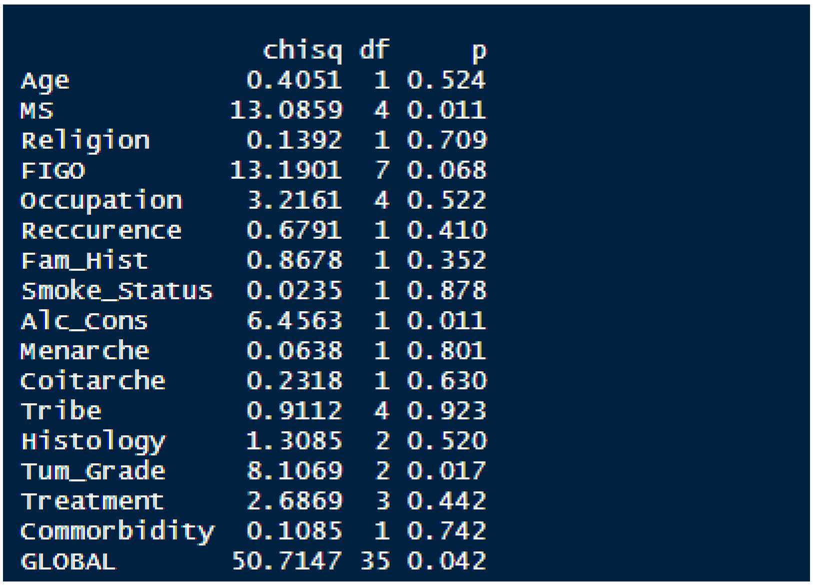
Testing the Cox Proportional Assumptions

The Cox proportional hazards model makes several assumptions. Thus, assessing whether a fitted Cox regression model adequately describes the data is important.

The three types of diagnostics for the Cox model are:

- Testing the proportional hazards assumption.
- Examining influential observations (or outliers).
- Detecting nonlinearity in the relationship between the log hazard and the covariates.

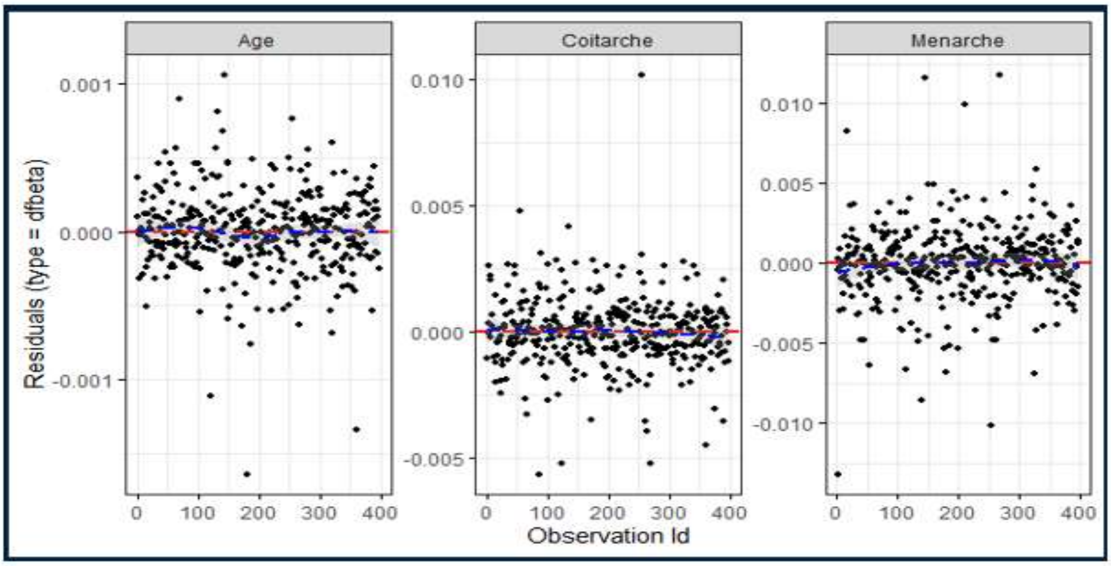
Testing the proportional hazards assumption.



```
                chisq df     p
Age            0.4051  1 0.524
MS            13.0859  4 0.011
Religion       0.1392  1 0.709
FIGO          13.1901  7 0.068
Occupation     3.2161  4 0.522
Reccurence     0.6791  1 0.410
Fam_Hist       0.8678  1 0.352
Smoke_Status   0.0235  1 0.878
Alc_Cons       6.4563  1 0.011
Menarche       0.0638  1 0.801
Coitarche      0.2318  1 0.630
Tribe          0.9112  4 0.923
Histology      1.3085  2 0.520
Tum_Grade      8.1069  2 0.017
Treatment      2.6869  3 0.442
Commorbidity   0.1085  1 0.742
GLOBAL        50.7147 35 0.042
```

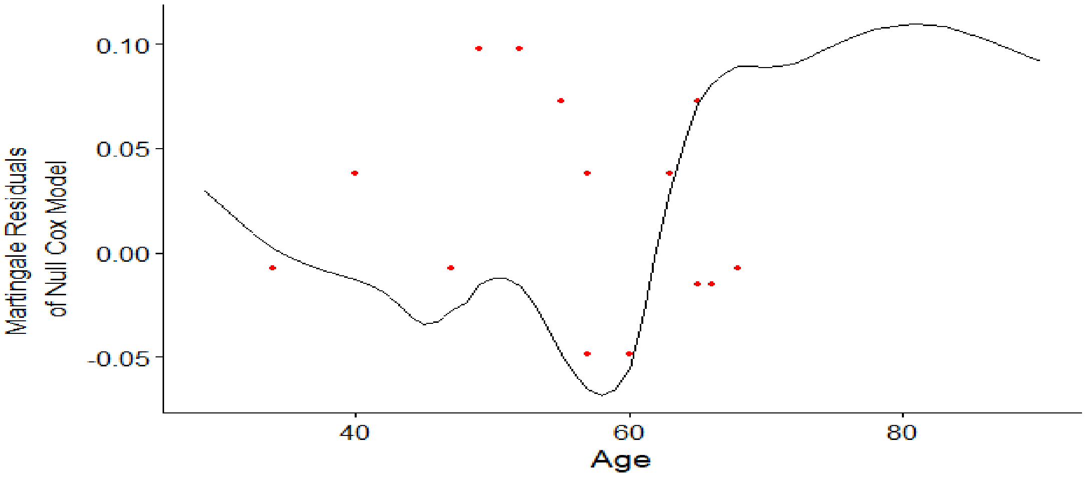
From the output above, the test is not statistically significant for some of the covariates but statistically significant for some, and the global test is also statistically significant at a 5% level of significance. Therefore, we cannot assume the proportional hazards**.**

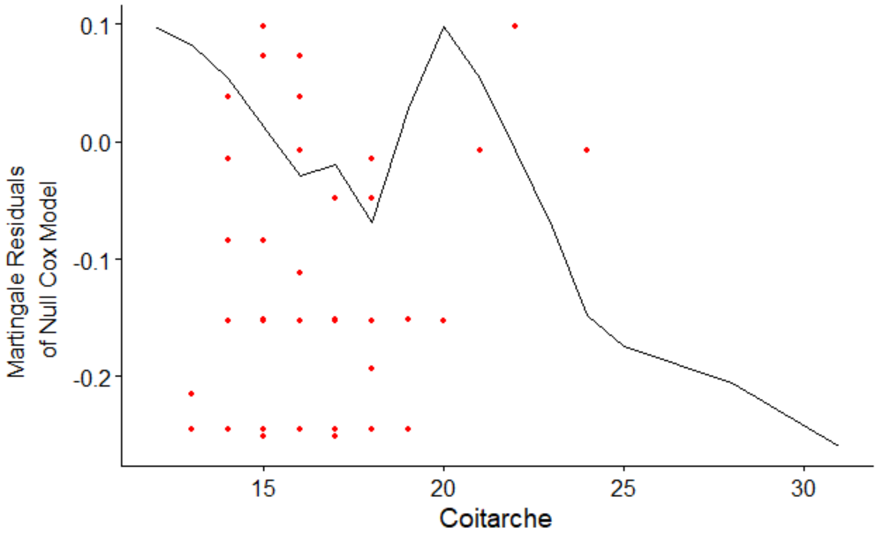
Examining influential observations (or outliers)

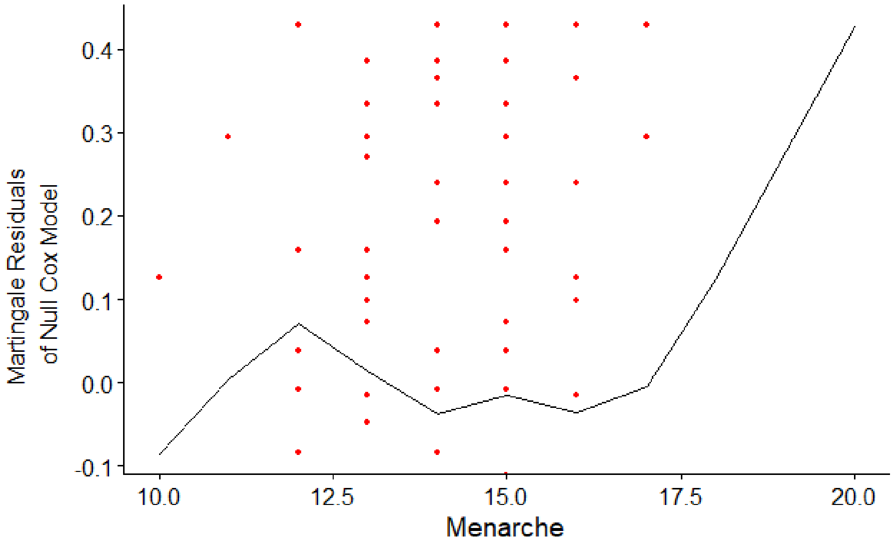


The above index plots show that comparing the magnitudes of the largest dfbeta values to the regression coefficients suggests that some of the observations are terribly influential individually, even though some of the dfbeta values for age, coitarche, and menarche are within the lines. Hence, outliers are observed in the data which makes cox PH model, not a good fit.

Detecting nonlinearity in the relationship between the log hazard and the covariates (continuous variables)







It appears that, nonlinearity has failed here.

APPENDIX V

R Codes

~~~~
```{r}
######### read the clinical data in#############
getwd()
setwd("C:/Users/USER/Documents/Fatima")
cervdata<-read.csv("Fat.csv")
```
```{r}
######### Run Kaplan-Meyer analysis #############
#Fit and plot Kaplan-Meyer curves:
kmfit <- survfit(Surv(dtime, Event) ~ 1, data = cervdata)
summary(kmfit)
library(ggplot2)
autoplot(kmfit,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Overal Survival Estimates of Patients with Cervical cancer")
# K-M for patients based on Marital Status
km_ms <- survfit(Surv(dtime, Event) ~ MS,
      data = cervdata)
autoplot(km_ms,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Marital Status")
# K-M Based on Religion
km_rel <- survfit(Surv(dtime, Event) ~ Religion,
      data = cervdata)
autoplot(km_rel,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer patients by Religion")
# K-M Based on FIGO
~~~~

```
km_figo <- survfit(Surv(dtime, Event) ~ FIGO,
        data = cervdata)
autoplot(km_figo,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Stage of Cancer")
# K-M Based on Occupation
km_occ <- survfit(Surv(dtime, Event) ~ Occupation,
        data = cervdata)
autoplot(km_occ,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimate of Cervival Cancer Patients by Occupation")
# K-M Based on Recurrence
km_rec <- survfit(Surv(dtime, Event) ~ Reccurence,
        data = cervdata)
autoplot(km_rec,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimate of Cervival Cancer Patients by Recurrence level")
# K-M Based on Family History
km_famhist <- survfit(Surv(dtime, Event) ~ Fam_Hist,
        data = cervdata)
autoplot(km_famhist,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Family History")
# K-M Based on Smoking Status
View(cervdata)
km_smokst <- survfit(Surv(dtime, Event) ~ Smoke_Status,
        data = cervdata)
autoplot(km_smokst,  xlab = "Time in months",
    ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Smoking Status")
# K-M Based on Alcohol
km_alc <- survfit(Surv(dtime, Event) ~ Alc_Cons,
        data = cervdata)
autoplot(km_alc,  xlab = "Time in months",
     ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Alcohol Consumption")
ggsurvplot(km_alc, data = cervdata)
# K-M Based on Tribe
str(cervdata$Tribe)
tribe <- as.factor(cervdata$Tribe)
km_tribe <- survfit(Surv(dtime, Event) ~ tribe,
        data = cervdata)
ggsurvplot(km_tribe, data = cervdata,  xlab = "Time in months",
     ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Tribe")
# K-M Based on Histology
km_hist <- survfit(Surv(dtime, Event) ~ Histology,
        data = cervdata)
autoplot(km_hist,  xlab = "Time in months",
     ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Histology")
# K-M Based on Tumour Grade
km_grade <- survfit(Surv(dtime, Event) ~ Tum_Grade,
        data = cervdata)
autoplot(km_grade,  xlab = "Time in months",
     ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Tumour Grade")
# K-M Based on Treatment
km_trt <- survfit(Surv(dtime, Event) ~ Treatment,
        data = cervdata)
autoplot(km_trt,  xlab = "Time in months",
     ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Treatment Types")
# K-M Based on Comorbidity
```

~~~
km_comor <- survfit(Surv(dtime, Event) ~ Commorbidity,
        data = cervdata)
autoplot(km_comor,  xlab = "Time in months",
     ylab = "Survival Probability", main = "Survival Estimates of Cervical Cancer Patients by Commorbidity")
```
```{r}
######### Run Log-rank test #############
# Find and plot log- ranks test for the two groups of the tumors
ms_dif <- survdiff(Surv(dtime, Event) ~ MS,
        data = cervdata)
ms_dif
#define a custom function to pull the p-value out of the Log-rank test
getPval <- function(x) {
if( is.matrix(x$obs))
etmp <- apply(x$exp, 1, sum)
else
etmp <- x$exp
df<- (sum(1 * (etmp > 0))) - 1
pv <- 1 - pchisq(x$chisq, df)
format(pv, digits = 3)
}
pValue<-getPval(ms_dif)
pValue
legend("bottomright", paste("p-value=", pValue), col = "black")
```
```{r}
##remove the cases with missing clinical variables
cervdata2<-na.omit(cervdata2)
dim(cervdata2)
head(cervdata2)
~~~

~~~
# Run coxph to fits a Cox proportional hazards regression model

coxfit <- coxph(formula = Surv(time, event) ~ Ms +Rel + Figo + Occ + Rec + Famhist
+ Smokst + Alc + Trbe + Hist + Grade + Trt + Comor + age + parity + menopause +
age_Diag + age_1st_Birth + age_last_Birth + menarche + coitarche , data = cervdata2,
ties = 'breslow')

summary(coxfit)
~~~

Appendix C

~~~
######### Run AFT model with specific distribution #############

# Apply AFT with four distributions

```{r}

expo = survreg(formula = Surv(dtime, Event) ~ Age + ms + rel + Parity + figo + occ +
Age_1st_Birth + Age_last_Birth + rec + famhist + smokst + alc + Menarche +
Coitarche + tribe + hist + grade + trt + comor,  data = cervdata , dist = "exponential")

aicexpo = extractAIC(expo)

summary(expo)

```

```{r}

log = survreg(formula = Surv(dtime, Event) ~ Age + ms + rel + Parity + figo + occ +
Age_1st_Birth + Age_last_Birth + rec + famhist + smokst + alc + Menarche +
Coitarche + tribe + hist + grade + trt + comor,  data = cervdata , dist = "loglogistic")

aiclog = extractAIC(log)

summary(log)

```

```{r}

logn = survreg(formula = Surv(dtime, Event) ~ Age + ms + rel + Parity + figo + occ +
Age_1st_Birth + Age_last_Birth + rec + famhist + smokst + alc + Menarche +
Coitarche + tribe + hist + grade + trt + comor,  data = cervdata , dist = "lognormal")

aiclogn = extractAIC(logn)

summary(logn)

```

```{r}

wei = survreg(formula = Surv(dtime, Event) ~ Age + ms + rel + Parity + figo + occ +
rec + famhist + smokst + alc + Menarche + Coitarche + tribe + hist + grade + trt +
comor,  data = cervdata , dist = "weibull")
~~~

~~~
aicwei = extractAIC(wei)
summary(wei)
```
```{r}
aicexpo
aiclog
aiclogn
aicwei
```
```{r}
library(AICcmodavg)
models = list(expo, log, logn, wei)
mod.names = c("exponential", "loglogistic", "lognormal", "weibull" )
aictab(cand.set = models, modnames = mod.names)
```
```
summary(weibul12)
hat.sig = weibul12$scale
hat.alpha = 1/hat.sig
reg.linear = weibul12$linear.predictor
reg.linear.mdf = -reg.linear/hat.sig
tt=cbind(Surv(clinData3$BCR_FreeTime, clinData3$BCR_Event))[,1]
cs.resid = exp(reg.linear.mdf)*tt^(hat.alpha)
cs.fit = survfit(Surv(cs.resid,clinData3$BCR_Event)~1,type="fleming-harrington")
summary(cs.fit)
par(mfrow=c(1,1))
plot(cs.fit$time, -log(cs.fit$surv),type = 's',xlab="Cox-Snell residual",ylab="Cumulative hazard of residual",main="Cox-snell plot for weibull model")
abline(0, 1, col = 'red', lty = 2)
```
Call:
~~~

~~~
survreg(formula = Surv(dtime, Event) ~ Age + ms + rel + Parity +
    figo + occ + Age_1st_Birth + Age_last_Birth + rec + famhist +
    smokst + alc + Menarche + Coitarche + tribe + hist + grade +
    trt + comor, data = cervdata, dist = "lognormal")
```{r}
## Testing the assumptions of Cox PH Model
install.packages(c("survival", "survminer"))
library("survival")
library("survminer")
## 1. Testing the proportional hazards assumption
## 2. Examining influential observations (or outliers).
## 3. Detecting nonlinearity in relationship between the log hazard and the covariates.
#########Computing a Cox Model############
View(cervdata)
res.cox <- coxph(Surv(dtime, Event) ~ Age+MS+Religion+FIGO+Occupation+Reccurence+Fam_Hist+Smoke_Status+Alc_Cons+Menarche+Coitarche+Tribe+Histology+Tum_Grade+Treatment+Commorbidity, data =  cervdata)
res.cox
res.cox2 <- coxph(Surv(dtime, Event) ~ Age + Menarche + Coitarche, data =  cervdata)
res.cox2
```

```{r}
test.ph <- cox.zph(res.cox)
test.ph
test.ph2 <- cox.zph(res.cox2)
test.ph2
```

```{r}
ggcoxzph(test.ph)
ggcoxzph(test.ph2)
```
~~~

~~~~
```{r}
ggcoxdiagnostics(res.cox, type = "dfbeta",
              linear.predictions = FALSE, ggtheme = theme_bw())
ggcoxdiagnostics(res.cox2, type = "dfbeta",
              linear.predictions = FALSE, ggtheme = theme_bw())
```
```{r}
ggcoxfunctional(Surv(dtime, Event) ~ Coitarche, data = cervdata)
```
~~~~