PROGNOSTIC FACTORS OF SURVIVAL TIME FOR BREAST CANCER PATIENTS IN PIETERSBURG HOSPITAL IN LIMPOPO PROVINCE OF SOUTH AFRICA

by

COLLEN HLUNGWANI

DISSERTATION

Submitted in fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematics and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Dr KD Moloi

CO-SUPERVISOR: Dr F Ooko

2024

Declaration

I, Collen Hlungwani, the undersigned, hereby declare that the work contained in this dissertation submitted to the University of Limpopo, for the qualification of Master of Science in Statistics is my own independent work and has not been submitted by me for a qualification at this or any other institution of higher learning. In addition, I confirm that all the references materials contained therein have been acknowledged to the fullest.

Signature:  _____ **Date: 09/12/2024.**

Hlungwani, C (Mr)

Abstract

The breast cancer metastasis in women had devastating disease throughout the world with patient goals aimed at improving quality of life and prolonged survival. Advanced treatments have been impressive for women with early-stage disease, and women diagnosed with advanced forms of breast cancer have also benefitted. To achieve this goal, more information is needed. Hence, the purpose of the study was to apply survival models and methods to identify breast cancer-related survival factors. These models were exponential, Weibull, log-normal, log-logistic, Cox proportional hazard and Kaplan-Meier method. The findings revealed that pre-menopause, human epidemic growth factor 2, positive oestrogen receptor, positive progesterone receptor, AJCC stage 4, endocrine therapy, surgery, N-stage, M1 stage, T-stage were significantly associated with better prognosis in Petersburg Oncology Hospital. It was further concluded that breast cancer patients who experienced lower survival are those breast cancer women who were diagnosed at late-stage when cancer cells had already advanced.

Dedication

I would like to dedicate this dissertation to my wife, Ascentia Mashumu, my children and my mother, Kwatisa Nwa-Masiya Hlungwani for their support and prayers.

Acknowledgements

I would like to thank my supervisor Dr K.D Moloi for his commitment, dedication and patience in supervising me. I also like to thank Dr F Ooko, my co supervisor for his support in all sessions and meetings we have had and for providing me with breast cancer data which I used in this study. Furthermore, I would like to thank the University of Limpopo for providing me with financial assistance to further my studies.

Contents

List of Figures

List of Tables

List of Special Symbols

θ	is a vector of regression coefficients/unknown parameter/ denotes all model parameters
S	generic symbol for a survival function
X_i	is a vector of explanatory variables
x_i	a vector containing the covariate information for subject i,
Y_i	is a vector of explanatory variables
T_i	observed event time
δ_i	is the event indicator
∂	partial symbol is used to represent a partial differential
F	generic symbol for a cumulative density function
N	normal distributed random variable
L	generic symbol for likelihood function
W	Weibull distributed random variable
f	generic symbol for a probability density function
$\hat{f}$	is an estimator of the density f
G	generic symbol for a cumulative density function
Δ	is a delta symbolising censoring indicator
δ	is the scale parameter
S_x	denote the survival function in a reference group (group x)
S_y	denote the survival function in a reference group (group y)
μ	is the intercept

∞	infinity
$S_l(t)$	survival distributor
$\hat{S}_L(t)$	estimated survival distributor
$\bar{x}_p$	an estimate of the median
x_p	a confidence interval for x_p
$\hat{S}(x_p)$	estimated survival of confidence interval
$w(t)$	weight function
O_j	corresponds to the observed number of events in the first group
E_j	to the expected number of events in the first group
$\hat{h}(t)$	cumulative hazard as the estimate
$\widehat{H}(t)$	is any estimator of the cumulative hazard rate
$h_1(t)$	are hazard functions of populations
$K(.)$	is a Kernel density function.
$\hat{S}(t)$	is the estimated probability that an individual will survive past time t
d_j	will denote the number who die at this time
n_j	individuals who are alive
$\hat{p}_j$	is the estimated survival probability
l_i	the truncation time
y_i	event/censored time of subject
V	variance
$\hat{V}$	estimated variance
α	is a vector containing the parameters
$\boldsymbol{\beta}$	is the vector of coefficients of each covariate

β_0	is a vector containing the hypothesized values
$\hat{\beta}$	the maximum likelihood estimator
$r_{s_{ik}}$	is the partial residuals
$r_{sc_{ik}}$	is the score partial residuals
$\in$	“is an element of” symbol
r_{mci}	The modified Cox-Snell residual designed to handle right-censored data
λ	generic symbol for a hazard function
$\hat{\lambda}$	is the estimated hazard function
t_p	the specific time point at which the survival probability is being evaluated
$\hat{t}_p$	the estimated specific time point at which the survival probability is being evaluated
Ψ	constant hazard function
$\widehat{\Psi}$	this is the estimated hazard ratio of the second versus the first group
π	this is pi and is approximately equal to 3.14159
$\hat{\zeta}$	sigma variant
ρ	shape parameter
Φ	is called the parameter acceleration factor
Γ	Gamma function
γ	Lower gamma function

List of Acronyms

AFT	Accelerated Failure Time
AIC	Akaike Information Criterion
AJCC	American Joint Committee on Cancer
ARID1B	AT-Rich Interaction Domain 1B
BC	Breast Cancer
BMI	Body Mass Index
BRCA1	Breast Cancer Gene 1
BRCA2	Breast Cancer Gene 2
CDKN1B	Cyclin-Dependent Kinase Inhibitor 1B
CK5/6	Cytokeratin 5/6
CPH	Cox proportional hazard model
CTCA	Cancer Treatment Centres of America
DNA	Deoxyribonucleic Acid
OR	Oestrogen Receptor
HDI	human development index
HER2	Human epidermal growth factor receptor 2
HIV	Human Immunodeficiency Virus
HR	Hazard Ratios
HSES	Household Socioeconomic Status
MAP3K1	Mitogen-Activated Protein Kinase Kinase 1
MBC	Metastasis Breast Cancer

MRI	Magnetic Resonance Imaging
NAC	Neoadjuvant Chemotherapy
NCOR1	Nuclear Receptor Corepressor 1
PR	Progesterone Receptor
SMARCD1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily D member 1
TNBC	Triple-Negative Breast Cancer
TNM	Tumour, Node, Metastasis

Chapter 1

Introduction

1.1 Background of the study

One of the major causes of mortality worldwide is cancer (Momenimovahed and Salehiniya, 2017). Incidence and mortality from breast cancer (BC) in Southern Africa by age (premenopausal and postmenopausal) in 2020 are 19.2 per 100,000 and 58.5 per 100,000, respectively (Arnold *et al.,* 2022). According to Fernandez-Moya et al. (2020), approximately 90% of BC tumors may be triggered by somatic driver mutations that initiate the carcinogenic process. The majority of identified BC driver genes are ARID1B, CASP8, MAP3K1, MAP3K13, NCOR1, SMARCD1, CDKN1B, AKT2, and TBX3. In addition, BC gens BRCA1 and BRCA2 are tumour suppressor genes called pathogenic mutations and when they change, BC can develop. According to Elmore *et al.,* (2021), BC was the second highest dominant cancer among Zimbabwean women patients. Based on the findings of Karutjaiva (2021), BC was the primary cause of many deaths in Namibia. Bhuiyan *et al.,* (2022) In South Africa lack of knowledge about BC and its symptoms, which were deemed not sore and not really concerning, was the primary factor in the detection of late-stage disease. Hence, the BC survival rate was very poor (Ayeni *et al.,* 2023). Almost everywhere in the globe, cancer is becoming more common, although, in many

wealthy nations, the mortality rate from BC is dropping (Ji *et al*., 2020) and increasing in Sub-Saharan countries. Sinha et al. (2022) found that women diagnosed with advanced-stage BC were more likely to be from sub-Saharan countries, with prevalence rates of 64.7% in Botswana and 63.3% in South Africa, compared to 13% in the United States.

Women younger than 45 years at diagnosis had lower survival point estimates In Sub-Saharan countries (Joko-Fru *et al.,* 2020). Most literature have indicated in Sub-Saharan countries stage at diagnosis for BC is typically late rendering the treatment less likely to be successful as compared to advanced countries (Scheel *et al.,*2018). An extensive review of the BC literature revealed no studies conducted in Limpopo Province, if any exist, that specifically investigate the predictors of survival in BC patients. Therefore, this study will employ proportional hazard and accelerated failure time (AFT) models to examine the relationship between BC patient survival and common prognostic factors, including age, disease stage, estrogen receptor status, progesterone receptor status, HER2 receptor status, and menopausal status.

1.2 Rationale

Despite the widespread of BC disease, there are notable regional differences in incidence, mortality, and survival rates. Genetics, lifestyle, population structure, and environment may be some of the causes of these variances, among other differences (Hortobagyi *et al.,* 2005). Parkin and Fernández, (2006) indicated that changes in risk variables have increased the prevalence of BC, which rises daily. According to Mavaddat *et al.,* (2015), even though the burden of BC can be lowered by screening people, the drawbacks of this approach include side effects, over diagnosis, and higher expenses. According to Mavaddat *et al.,* (2015), stratifying women based on their BC risk factors can be useful for enhancing risk-free practices and creating customized BC screening programs. According to Simms *et al.,* (2013),

prognostic factors are used by medical practitioners to make informed decisions on the timing of starting, changing, stopping or selecting a patient's particular therapy.

The Cox proportional hazard model (Cox, 1972) is one of the most commonly used models in survival analysis. The Cox proportional hazard model (CPH) model's time-to-event prediction is not trivial, hence, Breslow's estimator is a non-parametric estimator that must be used to estimate the baseline hazard function, in order to apply the CPH model (Basha and Hoxha, 2019). Breslow's estimator computation is computationally demanding for massive data and requires access to the training data.

Although less popular than Cox proportional hazard the Accelerated failure time model is another well-known survival analysis technique. There are two main reasons why we decided to investigate Accelerated failure time in our study. First, the predictive analysis in addition to analyzing model parameters would be taken as coefficients. Second, if the Cox proportional hazard assumption is false, the Accelerated failure time model might offer a better fit (Faruk, 2018).

1.3 Aim and objective of the study

1.3.1 Aim

The aim of the study is to identify prognostic factors that contribute to breast cancer in Pietersburg Oncology Hospital

1.3.2 Objectives

The objectives of the study are to:

i. Compare the performance of proportional hazard and accelerated failure time models.
ii. Assess the performance of prognostic model.
iii. Assess the performance of predictive models.
iv. Stratify BC patients into higher or lower risk categories.
v. Identify potential clinical prognostic factors.

1.4 Scientific Contribution

A substantial number of women are more likely to be diagnosed with aggressive or advanced forms of breast cancer. Black women seem to be more likely to die from breast cancer than women of all other racial and ethnic groups. These disparities or inequalities are thought to reflect the interplay of many prognostic factors, from tumour biology to matters like income, diet, access to quality health care, and other factors.

There exists a lot of confusion about when women should be assessed for their risk of breast cancer and when should begin for the disease. The existing guidelines from medical groups vary in the age at which they recommend beginning screening: some are advised at the age of 50, while others recommend beginning at the age of 40 or 45.

The study intends to contribute to the scientific discussion through its findings especially by addressing the critical gap in breast cancer risk prediction using predictive models and prognostic models. These risk models are intended to also help inform the discussion that is presently going on. The models could also be used

to identify women that are eligible to participate in studies to prevent breast cancer in women at high risk of the disease.

1.5 Structure of the Dissertation

The study consists of five chapters. Chapter 1 presents the introduction, problem statement, rational, aim and objective of the study, scientific contribution and structure of the study. Chapter 2 reviews literature of factors that led to breast cancer in women in South Africa and other relatable parts of the world. Chapter 3 presents the statistical analysis methods that were used in the study. Chapter 4 presents the results and discussion of the study. Chapter 5 presents the conclusion, recommendation, and limitation of the study based on the findings generated in this study.

Chapter 2

Literature review

2.1 Introduction

This chapter reviews materials such as articles, journals, manuscripts, books, and any other materials relevant to the study of breast cancer. In the conclusion of the chapter, findings from several research studies are explored, summarised and links between the researcher's works are established.

2.2 Literature reviewed

Cancer is one of the major causes of mortality worldwide (Momenimovahed and Salehiniya, 2017). Saini *et al.,* (2020) define Cancer as a genetic disorder that results from genetic or epigenetic alterations in the somatic cells and has abnormal cell growth which may be spread to other body parts. According to Sung *et al.,* (2020) globally, approximately 19.3 million new cancer cases were reported in 2020, excluding nonmelanoma skin cancer, which accounted for 18.1 million cases. During the same year, cancer-related deaths reached nearly 10.0 million, or 9.9 million when excluding nonmelanoma skin cancer. Female breast cancer has become the most frequently diagnosed cancer, overtaking lung cancer, with an estimated 2.3 million new cases (11.7%). It is followed by lung cancer (11.4%), colorectal cancer (10.0%), prostate cancer (7.3%), and stomach cancer (5.6%) (Sung *et al*., 2020).

Fina (2022) described breast cancer as a disease characterized by abnormal cell growth that can spread outside the breast through blood vessels and lymph vessels. Breast cancer is classified into three main types based on the presence or absence of specific proteins in cancer cells. Hormone receptor-positive breast cancer, accounting for 70% of cases, is characterized by the presence of oestrogen receptor (ER) or progesterone receptor (PR) proteins. ERBB2-positive breast cancer (previously referred to as HER2-positive) represents 15% to 20% of cases and is marked by elevated levels of the ERBB2 protein. Triple-negative breast cancer, comprising 15% of cases, lacks ER, PR, and ERBB2 proteins in the cancer cells. (Waks and Winer, (2019)).

A considerable amount of literature on prognostic factors of breast cancer has been published across the world. However, there has been relatively little literature published in Limpopo Province, to investigate the predictors of survival in patients with BC.

Momenimovahed and Salehiniya, (2019) revealed the risk factors of breast cancer in the world. Included among the factors were hereditary factors, lifestyle, hormonal and demographics. Hereditary factors refer to family history and genetic factors whilst lifestyle refers to obesity and overweight, excess consumption of alcohol, smoking and diet. Furthermore, hormonal factors refer to long-term use of contraceptives or postmenopausal hormone therapy. Another factor is demographic such as gender which is unique to women and is a rare malignancy in men, accounting for less than 1% of all cases of breast cancer. Moreover, age was indicated to be another demographic factor of breast cancer that increased significantly with age and reaches its peak in the age of menopause and then gradually decreases or remains constant.

A study by Arnold *et al.,* (2022) demonstrated that In 2020, the incidence and mortality rates of breast cancer (BC) in Southern Africa were 19.2 per 100,000 and

58.5 per 100,000, respectively, with distinctions observed between premenopausal and postmenopausal age groups. Fernandez-Moya *et al*., (2020) reported that approximately 90% of BC tumours may result from somatic driver mutations initiating the carcinogenic process. Key BC driver genes include ARID1B, CASP8, MAP3K1, MAP3K13, NCOR1, SMARCD1, CDKN1B, AKT2, and TBX3. Furthermore, the tumour suppressor genes BRCA1 and BRCA2, known as pathogenic mutations, play a critical role in BC development when altered. Elmore *et al*., (2021) highlighted BC as the second most prevalent cancer in Zimbabwe, while Karutjaiva (2021) identified it as the leading cause of death in Namibia.

In addition Bhuiyan *et al*., (2022) noted that a significant factor contributing to the late-stage diagnosis of breast cancer (BC) in South Africa is a general lack of awareness about the disease and its symptoms, which are often perceived as non-painful and therefore not alarming. Consequently, the survival rate for BC in South Africa remains notably low (Ayeni *et al*., 2023). Globally, cancer incidence is rising, but while BC mortality rates are declining in many high-income countries, they are increasing in Sub-Saharan Africa (Ji *et al*., 2020). Siegel *et al*, 2018 agrees with Ji *et al.,* (2020) that between 1991 and 2018, the cancer mortality rate decreased steadily by 31%. This decline is attributed to reductions in smoking, as well as advancements in early detection and treatment methods. Sinha *et al*., (2022) found that women diagnosed with advanced-stage BC were more likely to be from sub-Saharan countries, with prevalence rates of 64.7% in Botswana and 63.3% in South Africa, compared to 13% in the United States. In Francies *et al*., (2020) developed nations like the US are reported to have high breast cancer incidence and mortality rates. In contrast, developing countries exhibit lower incidence rates but face rising mortality, indicating a lack of resources for preventive screening, early detection, and effective treatment. Variations in incidence rates between countries are influenced by differences in exposure to environmental risk factors, as well as the behaviours and lifestyles of distinct population groups.

Women younger than 45 years at diagnosis had lower survival point estimates in Sub-Saharan countries (Joko-Fru *et al.,* 2020). Most literature have indicated in Sub-Saharan countries stage at diagnosis for BC is typically late rendering the treatment less likely to be successful as compared to advanced countries (Scheel *et al.,*2018).

A study was conducted to predict survival for women with invasive breast cancer on race (Walsh, *et al.,* 2019). The results indicated that black women with breast cancer have lower survival rates and higher recurrence rates in comparison with white women. Walsh, *et al.,* (2019) concluded that black women had more advanced disease and adverse prognostic indicators at diagnosis, but race was not an independent predictor of outcome. Black women were significantly more likely to have triple negative breast cancer.

In another study, Johansson *et al.,* (2019) found age and tumour subtype to be prognostic factors for breast cancer survival, but it is unclear which matters the most. Johansson *et al.,* (2019) used population-based data to address what matters most between age and tumour. A study conducted using data from the Cancer Registry of Norway identified 21,384 women aged 20–89 diagnosed with breast cancer between 2005 and 2015. Breast cancer subtypes were classified based on oestrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status: luminal A-like (ER+PR+HER2-), luminal B-like HER2-negative (ER+PR-HER2-), luminal B-like HER2-positive (ER+PR+/-HER2+), HER2-positive (ER-PR-HER2+), and triple-negative (TNBC) (ER-PR-HER2-). Using Cox regression analysis, hazard ratios (HR) 7-year breast cancer-specific survival were estimated by age and subtype, adjusting for year, grade, TNM stage, and treatment. The findings revealed that younger women were more likely to have HER2-positive and TNBC tumours, while women aged 70–89 were more likely to have luminal A-like tumours. Compared to women aged 50–59, younger women showed a twofold increase in breast cancer-specific mortality, whereas elderly women exhibited two to five times higher. After adjustments, the association among

younger women became non-significant, but mortality remained significantly elevated among the elderly. Young age was linked to higher breast cancer-specific mortality in the luminal A-like subtype, while advanced age was associated with higher mortality across all subtypes. Age and subtype emerged as strong independent prognostic factors, with elderly patients consistently showing worse outcomes, even after adjusting for subtype. Tumour characteristics, including subtype, grade, and stage, largely explained the higher mortality rates observed in younger women.

A research by Grimm *et al.,* (2022) was carried out to identify various methods to diagnose breast cancer which included breast ultrasound, fine needle aspiration, needle biopsy, punch biopsy, vacuum assisted biopsy, and wire guided excision biopsy, breast MRI scan and Mammogram. In addition, Grimm *et al*. (2022) examined the benefits and risks of mammography screening in women aged 40 to 49, highlighting that younger women often present with breast cancer at a more advanced stage. This includes larger tumour sizes, lymph node involvement, and more biologically aggressive profiles characterized by lower oestrogen receptor (ER) positivity, HER2 overexpression, and higher nuclear grades. As a result, younger women tend to have poorer prognoses, with a higher likelihood of recurrence and breast cancer-related mortality compared to older women. Additionally, breast cancer in this age group is more frequently associated with genetic mutations. Notably, premature deaths from breast cancer in women in their 40s result in as many years of life lost as those in their 50s and significantly more years lost than diagnoses occurring in women in their 60s.

Ayeni *et al.,* 2020 researched about the multimorbidity profile of South African women newly diagnosed with breast cancer. It was found that multimorbidity in women with breast cancer may delay presentation, and affect treatment decisions and outcomes. The description of the multimorbidity profile of women with breast

cancer, its determinants, associations with stage at diagnosis, and treatments was received by women with BC.

Ayeni *et al.,* 2020 collected self-reported data on five chronic conditions (hypertension, diabetes, cerebrovascular diseases, asthma/chronic obstructive pulmonary disease, tuberculosis), determined obesity using body mass index (BMI) and tested HIV status, in women newly diagnosed with breast cancer between January 2016 and April 2018 in five public hospitals in South Africa. The identification of determinants of greater than 2 of the seven above-mentioned conditions (defined as multimorbidity), multimorbidity itself with stage at diagnosis (advanced [III–IV] vs. early [0–II]) and multimorbidity with treatment modalities was made. Among 2,281 women, 1,001 (44%) presented with multimorbidity. Obesity (52.8%), hypertension (41.3%), HIV (22.0%) and diabetes (13.7%) were the chronic conditions that occurred most frequently. Multimorbidity was more common with older age (OR = 1.02; 95% CI 1.01–1.03) and higher household socioeconomic status (HSES) (OR = 1.06; 95% CI 1.00–1.13). Multimorbidity was not associated with advanced-stage breast cancer at diagnosis, but for self-reported hypertension there was less likelihood of being diagnosed with advanced-stage disease in the adjusted model (OR 0.80; 95% CI 0.64–0.98). Multimorbidity was associated with first treatment received in those with early-stage disease, p = 0.003. The prevalence of multimorbidity is high among patients with breast cancer. Our findings suggest that multimorbidity had a significant impact on treatment received in those with early-stage disease. There is need to understand the impact of multimorbidity on breast cancer outcomes.

A study was conducted by Kazmi *et al.,* (2020) to focus on breast cancer patients with other diseases. The study aimed to assess overall survival (OS) in real-world clinical settings for patients with metastatic breast cancer (MBC) and visceral metastases (liver or lung) treated with third-line therapies, including eribulin, gemcitabine, or capecitabine. The analysis focused on the overall population and major MBC subtypes: triple-negative breast cancer (TNBC), hormone receptor-

positive/human epidermal growth factor receptor 2-negative (HR+/HER2−), and HER2-positive (HER2+). Data were derived from de-identified electronic health records of patients at the Cancer Treatment Centers of America (CTCA). The study included patients diagnosed with MBC involving lung or liver metastases who received one of the specified third-line treatments. Landmark survival rates were determined at 6, 12, 24, and 36 months, and survival outcomes were compared between treatment groups within TNBC and HR+/HER2− subtypes using log-rank analysis. Cox regression analysis was employed to estimate hazard ratios for treatment comparisons within these subtypes. Results showed that 443 patients received third-line therapy: eribulin (n = 229), gemcitabine (n = 134), or capecitabine (n = 80). Patients treated with eribulin demonstrated higher survival rates at all landmark timepoints compared to gemcitabine and at 36 months compared to capecitabine. Median survival times indicated numerically superior overall survival for eribulin. Kazmi *et al*., (2020) concluded that the findings align with randomized clinical trial data, reinforcing the effectiveness of eribulin in treating MBC patients with lung or liver metastases, particularly in TNBC and HR+/HER2− subtypes.

Malik *et al.,* (2019) conducted a research on survival analysis of breast cancer patients with different treatments. The aim of the study was to explore and better understand clinic pathological details of breast cancer patients and analyse their survival rate among different treatment groups. To achieve the objectives data from September 2014 to February 2018 were collected from five hospitals in the region of Rawalpindi and Islamabad from Pakistan. The data comprised histo-pathologically confirmed breast cancer cases. Furthermore, Patient characteristics and medical history were collected using a detailed questionnaire. All the subjects were followed up, and information regarding their current health and treatment status was collected. The study included 347 participants with a mean age of 44.3 ± 12.2 years and an average body mass index (BMI) of 27.9 ± 4.0 kg/m². Key factors associated with breast cancer development included younger age, higher BMI, consanguinity, and family history ($p < 0.05$). Among the cases, invasive ductal carcinoma was the

most common subtype, affecting 267 individuals (77%), with Grade II tumours being most prevalent (234 cases, 67%). Positive lymph node involvement was observed in 221 cases (64%), and 97 patients (28%) had metastases to other organs.

Survival analysis demonstrated a statistically significant impact ($p < 0.0001$) of all treatment modalities on overall survival. Malik *et al*., (2019) concluded that combining various treatment approaches could lead to improved health outcomes for breast cancer patients.

Senyefia *et al.,* (2022) conducted a study to compare accelerated failure time (AFT) models in analysing the survival of breast cancer (BC) patients. The study highlighted that female breast cancer has now surpassed lung cancer as the leading cause of cancer-related diagnoses globally, with geographic variations in the impact of associated risk factors on survival outcomes. The objective of the study was to evaluate various survival models to determine the effect of risk factors on patient survival. The study analysed secondary data from 558 BC patients diagnosed at Korle Bu Teaching Hospital between 2010 and 2015, with follow-up data up to the end of 2015. The patients' survival status, along with demographic and tumour characteristics, were assessed using event history analysis. To compare the performance of different survival models, the study employed Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Receiver Operating Characteristic (ROC) curve analysis. R software was used for the data analysis. The data set consisted of patients aged 13 to 97 years, and was split into a training set (70%) and a validation set (30%). The results, based on AIC, BIC, and ROC curve values, indicated that the Gompertz model (AIC=2322, BIC=2391) was the best fit for the survival data. The Generalized Gamma (AIC=2378, BIC=2451) and Weibull (AIC=2382, BIC=2452) models were the next best alternatives among the nine AFT models considered. The three best-fitting AFT models revealed that significant covariates affecting survival included age at diagnosis, progesterone receptor status, molecular subtype, grade, stage, metastasis, number of lymph nodes

involved, and genetic status ($p<0.05$). The Gompertz model demonstrated excellent discriminatory power, with an area under the ROC curve (AUC) of 0.945, indicating outstanding performance in distinguishing the true disease status of patients.
In conclusion, while the Cox proportional hazards model has been widely used and remains a robust tool for survival analysis, its assumption of proportional hazards is often violated by some covariates in medical research. In such cases, AFT models provide a strong alternative

In the study titled Determinants of Survival Time of Women with Breast Cancer: Archives of Oncology and Cancer Therapy by Sharma and Abebe (2019), it was found that breast cancer, which originates from breast tissue, is the most commonly diagnosed cancer worldwide and has a high mortality rate. The objective of the study was to identify factors influencing the survival time of women with breast cancer. The study included 819 women with breast cancer, using data from medical records of patients enrolled between September 2016 and 2018 at Black Lion Hospital in Ethiopia. Kaplan-Meier plots and the Log-rank test were used to compare survival functions, while Cox Proportional Hazard (Cox-PH) and Accelerated Failure Time (AFT) models were employed to identify factors affecting survival. Among the participants, 39.60% died during the study period, and the median survival time was found to be 33 months. The results from the Weibull AFT model indicated that survival time was significantly influenced by factors such as age, use of oral contraceptives, alcohol consumption, breastfeeding, tumour size, histologic grade, and cancer stage. The Weibull AFT model provided a better fit for the breast cancer dataset compared to the other AFT models used in the study.

In a study conducted by Joko-Fru *et al*., (2020), the survival of breast cancer (BC) patients in Sub-Saharan Africa was estimated based on area, stage, and country-level Human Development Index (HDI). Despite the scarcity of population-level survival data from Africa, especially data on survival differences by stage at diagnosis, this study aimed to address this gap. Data was obtained from a random

sample of 2,588 BC cases diagnosed between 2008 and 2015 across 14 population-based cancer registries in 12 countries: Benin, Côte d'Ivoire, Ethiopia, Kenya, Mali, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Uganda, and Zimbabwe. Of these, 2,311 cases were included in the survival analyses. The 1-, 3-, and 5-year observed and relative survival (RS) rates were estimated by registry, stage, and country-level HDI, while excess hazards were calculated adjusting for potential confounders. Among patients with known stage at diagnosis, 64.9% were diagnosed at late stages, and 18.4% had metastatic cancer at diagnosis. RS varied significantly by registry, ranging from 21.6% (8.2–39.8) at Year 3 in Bulawayo to 84.5% (70.6–93.5) in Namibia. Patients diagnosed at early stages had a 3-year RS of 78% (71.6–83.3), compared to just 40.3% (34.9–45.7) for those diagnosed at advanced stages (III and IV). Overall, the RS was 86.1% (84.4–87.6) at Year 1, 65.8% (63.5–68.1) at Year 3, and 59.0% (56.3–61.6) at Year 5. Age at diagnosis was not independently associated with increased mortality risk after adjusting for stage and country-level HDI. The study concluded that improving breast cancer survival outcomes in Africa could be achieved by promoting early-stage diagnosis (down staging) and enhancing access to quality care.

A study by Bhuiyan *et al*., (2022) aimed to investigate factors contributing to late-stage presentation at the Mankweng Hospital Breast Cancer Clinic. It is well-documented that women in low- and middle-income countries often seek medical attention at advanced stages of breast cancer, which negatively impacts prognosis, regardless of the available healthcare resources. The study was conducted over 18 months, from July 2020 to December 2021, at the Mankweng Breast Cancer Clinic in the rural Limpopo province. A quasi-experimental design was used to assess the stage of breast cancer at the time of clinic presentation and the factors influencing late-stage diagnosis. The study sample consisted of 269 patients, of which 98% were female (n=269) and 2% were male (n=5). The majority of the women (75%, n=203) presented with late-stage cancer, while only 25% (n=66) were diagnosed at an early stage. Bhuiyan *et al*., (2022) concluded that 76% of the patients presented with late-

stage disease. The primary reason for the delay was a lack of knowledge about breast cancer and its symptoms, with many patients not recognizing the seriousness of the disease because it was not painful. A concerning finding of the study was that a higher proportion of educated patients presented with late-stage disease compared to those with a lower level of education.

In a study by Sakafu *et al.,* (2022), several factors contributing to delayed diagnostic presentation of breast cancer were identified. These included a lack of basic knowledge and awareness of breast cancer, stigma, financial barriers, and limitations within the local healthcare system. Additionally, the influence of friends and family played a significant role, acting as both facilitators and barriers to seeking timely medical attention. The study's findings aim to inform the development of educational intervention strategies that address these barriers and promote earlier diagnosis of symptomatic breast cancer in Tanzania.

Waks and Winer (2019) highlighted that most individuals diagnosed with breast cancer in the United States do not die from the disease. While approximately 250,000 cases are diagnosed annually, there are only about 41,000 breast cancer-related deaths each year. The prognosis and treatment of breast cancer depend on both the stage of the cancer and its specific type. The authors emphasized that treatment plans should be individualized, developed in collaboration between the patient and the oncologist. For patients with stage I to III breast cancer, the treatment goal is curative. This typically involves surgery to remove the tumour, medications, and potentially radiation therapy. In contrast, for patients with stage IV breast cancer, where the disease has spread to distant parts of the body, the primary goal is to control the cancer for as long as possible. Treatment for stage IV primarily involves medications. Waks and Winer (2019) also noted that different breast cancer types respond to specific treatments. Hormone receptor-positive breast cancer is often treated with anti-oestrogen medications, sometimes in combination with intravenous chemotherapy. For ERBB2-positive breast cancer, intravenous therapies targeting

the abnormal ERBB2 protein are used alongside chemotherapy. Finally, triple-negative breast cancer is treated with intravenous chemotherapy alone.

Sakafu *et al*., (2022) identified several common factors contributing to delayed breast cancer diagnosis in Tanzania, including a lack of basic knowledge and awareness, stigma, financial barriers, and local healthcare system challenges. The influence of friends and family was also significant, serving both as facilitators and barriers. These findings are essential for developing educational interventions aimed at improving the timely diagnosis of symptomatic breast cancer in the region.

Balhi (2023) further explored factors influencing breast cancer diagnosis and treatment in sub-Saharan Africa, highlighting that denial, the perception of breast cancer as an incurable disease, and the use of traditional medicine contributed to delays in seeking medical care. Studies by Gebremariam *et al*., (2019) and Ströbele *et al*., (2018) showed that reliance on traditional medicine before seeking medical consultation led to patient delay. Hassen *et al*., (2021) found that visiting traditional healers was a key factor in delayed presentation to health facilities. Additionally, Kohler *et al*., (2017) noted that patients often sought medical care only after trying multiple remedies, including herbal and spiritual treatments, such as holy water. This reliance on alternative medicine is further supported by a multi-country African study, where 71% of women believed in spiritual or faith healing, with belief varying by country (Balhi, 2023). Balhi (2023) also described common treatments for all breast cancer subtypes, including surgery, radiation therapy, and chemotherapy. Surgery options include mastectomy (total excision of the breast) followed by breast reconstruction or breast-conserving surgery (lumpectomy). Axillary lymph node removal is performed to determine cancer spread. Radiation therapy (RT) is important for reducing recurrence and mortality, especially in node-positive patients after mastectomy (Boyages, 2017). Chemotherapy, which involves cytotoxic drugs such as alkylating agents, antimetabolites, and tubulin inhibitors, is commonly used in both neoadjuvant (pre-surgery) and adjuvant (post-surgery) settings. Neoadjuvant

chemotherapy (NAC) is increasingly used for down staging the tumour before surgery, particularly in non-metastatic but inoperable breast cancer (Wang and Mao, 2020). The standard chemotherapy regimens often combine anthracyclines and taxanes, such as cyclophosphamide and doxorubicin followed by paclitaxel or docetaxel (Citron *et al*., 2003; Sparano *et al*., 2008)..

Chapter 3

Methodology and analytical procedure

3.1 Introduction

This chapter presents a review of the Cox proportional hazard model, Kaplan-Meier, and accelerated failure time models with special emphasis on exponential, Weibull, log-logistic, and log-normal models. The review will include the model framework, fitting, and model checking.

3.2 Data Collection

In this study, 333 patients with BC tumour characteristics and demographic will be studied. The data were obtained from patients diagnosed with BC from 01 January 2010 to 31 December 2020 at Pietersburg Oncology Hospital in Limpopo Province. The data will be collected from patients' information in the archives of the hospital which contains their survival status and treatment information over a period of time.

Only patients with complete information over the 10 year period will be selected for the study. Patients who did not experience the event (death due to BC) will be right-censored. The variables such as age at diagnosis, disease relapse status, progesterone receptor (PR) status, human epidermal growth factor 2 (HER2), receptor status, molecular subtype, oestrogen receptor (OR) status, disease stage (I, II, III, and IV), distance metastases, menopause status at diagnosis, will be analysed. The data will be partitioned into training (holding 70%) and validation set (30%).

Survival analysis is the study of survival times and factors that influence them. Types of studies with survival outcomes Include clinical trials, prospective and retrospective observational studies, and animals' experiments. Examples of survival times include time from birth until death, time from entry into a chemical trial until death or disease progression, or time from death, birth to development of breast cancer. The survival endpoint can also be referred to as positive event. For example, one might be interested in the time from entering into clinical trial until tumour response. Survival studies can include estimation of the survival distribution, comparison of the survival distributions of various treatments or interventions, or inclination of the factors that influence survival times.

In academic realism, survival analysis is now widely employed in a long list of applied sciences, owing considerably to the availability of longitudinal data that records histories of various survival processes and the occurrences of various events. At present, the concept of survival no longer simply refers to biomedical or demographic events; rather, it expands to indicate a much broader scope of phenomena characterized by the time-to-event process.

In medical research, clinical trials are regularly used to assess the effectiveness of new medicines or treatment of a disease. In this setting, researchers employ survival analysis to compare the risk of death or recovery from disease between or among

population groups receiving different medication or treatment. The result of such an analysis can provide important information with policy implication. Survival analysis is also applied in biological research. Mathematical biologists have long been interested in evolutionary perspective of human populations and their species.

By using survival analysis as the underlying means, they delineate the life history for a species population and link its survival process to a collection of physical attributes and behavioural characteristics for examining its responses to its environment. In social science, survival data are commonly collected and analysed with topics ranging widely from unemployment to drug use, recidivism, material disruption, occupational careers, and other social processes. In demography, in addition to the mortality analysis, researchers are concerned with such survival processes as the initiation of contraception use, internal and international migration, and the first Live birth intervals.

In the field of public health, survival analysis can be applied to the analysis of healthcare utilization. Such examination is of special importance for both planners and academics because the health services system reflects the political and economic organization of society and is concerned with fundamental issues involving life, death, and quality of life.

Survival analysis has a wide application in some other disciplines such as engineering, political science, business management, and economics. For example, in engineering, scientists apply analysis to perform life tests on the durability of mechanical or electrical products, the result of such studies can be used for the quality improvement of the product.

3.3 Censoring

Censoring is what distinguishes survival analysis from other fields of statistics. Basically, a censored observation contains only partial information about the variable of interest.

There are different types of censoring, here we consider type 1 right censoring only.

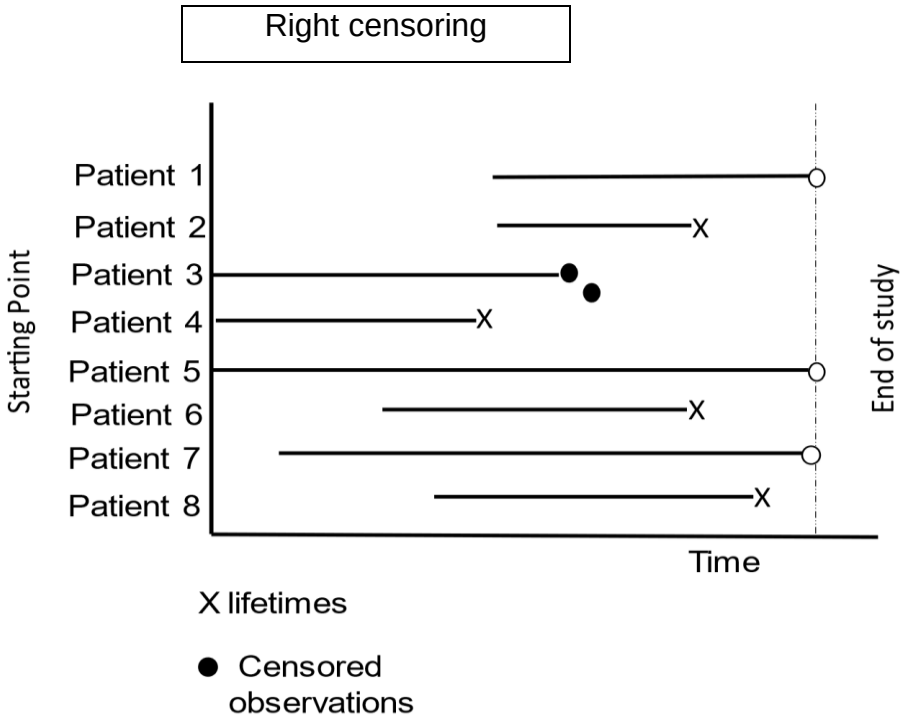


Figure 3. 1: Right Censored lifetimes of patients in artificial clinical trials

Let $X_1, X_2, X_3, \ldots, X_P$ be i.i.d survival times with cumulative distribution function F and let $Y_1, Y_2, Y_3, \ldots, Y_p$ be i.i.d. censoring times with cumulative distribution function G.

Throughout the (dissertation) study we assume that F and G are absolutely continuous. Furthermore, let f and g be probability density functions with respect to F and G.

We can only observe.

$$(T_1, \Delta_1), (T_2, \Delta_2), \dots, (T_p, \Delta_p) \quad \text{where}$$

$$T_i = min\{X_i, T_i\} \ and$$

$$\Delta_i = \begin{cases} 1 & : \ if \ \ X_i \leq Y_i, \ \ that \ is, \ \ T_i \ is \ not \ censored \\ 0 & : \ if \ \ Y_i \leq X_i, \ \ that \ is, \ \ \ T_i \ is \ censored \end{cases} \tag{3.1}$$

Random censored arises especially in medical applications, for example in clinical trials or epidemiological studies. Here, patients may enter the study at different times, then each is treated with one of several possible drugs or therapists. We are interested in observing their lifetimes, but censoring occurs in one of the following forms.

- Loss to follow up. The patient may move elsewhere; he/she is never seen again.
- Drop out. The treatment may have such strong side effects that is necessary to stop therapy. Or the patient may refuse to continue the treatment.
- Termination of the study. The study ends at a predefined point in time. This type of study is called administrative censoring.

We use X and Y, with no subscript, as shorthand for all the X_i and Y_i variables.

Assumption: Lifetimes X and censoring time Y are independent. A weaker condition is to assume that censoring is non-informative.

Remarks: The cumulative distribution function of the non-censored observations (while discarding the censored observations of the sample) is not F!

$$P(T < t\,, \Delta = 1) = P(X < t\,, X \leq y)$$

$$= \int_{x<t} \int_{x\leq y} f(x)\, g(y) dx dy$$

$$= \int_{x\leq y} f(x) \left(\int_{x\leq y} g(y) dy \right) dx$$

$$= \int_{x<t} f(x)(1 - G(x)) dx \quad \neq F(t)$$

(3.2)

Theorem 3.1. (Wienke (1996)) The probability density function of the data $((T, \Delta)$takes the form

$$f(t, \delta) = [f(t)(1 - G(t)]^{\delta} [g(t)(1 - F(t)]^{1-\delta} \tag{3.3}$$

Proof

Denote by H_0 and H_1 sub-distributions and each by h_0 and h_1, sub-densities. It holds by (3.1) that

$$H_1(t) = P(T < t\,, \Delta = 1) = \int_{x<t} f(x)[1 - G(x)] dx$$

Furthermore

$$H_0(t) = P(T < t\,, \Delta = 0)$$

$$=P(y < t\,.Y < X)$$

$$= \int_{y<t,} \int_{y\leq x} f(x)\, g(y) dxdy$$

$$= \int_{y<t} g(y) \left(\int_{y<x} f(x) dx \right) dy$$

$$= \int_{y<t} G(y) dy \big(1 - F(y)\big) dy$$

(3.4)

Consequently,

$$h_0(t) = H_0'(t) = g(t)[1 - F(t)]$$

$$h_1(t) = H_1'(t) = F(t)[1 - G(t)]$$

$$f(t, \delta) = \sigma h_1(t) + (1 - \delta) h_0(t)$$

$$= [h_1(t)]^{\delta} [h_0(t)]^{1-\delta}$$

$$= f(t)[1 - G(t)]^{\delta} [g(t)(1 - F(t)]^{1-\delta}$$

That completes the proof

Remark: If the censoring is non-informative, meaning that if the censoring distribution does not contain any information about the parameters of interest, then it does not enter the likelihood function:

$$L(t, \delta) = [f(t)]^{\delta} [1 - F(t)]^{1-\delta}$$

$$= \delta f(t) + (1-\delta)[1 - F(t)] \quad (3.5)$$

as it was pointed out in Theorem (3.1), the density function under independent right censoring is

$$f(t,\delta) = \delta f(t)[1 - G(t)] + (1-\delta)g(t)[1 - F(t)] \quad (3.6)$$

The following example considers the case of dependent censoring. It turns out that the likelihood function under censoring is a composition of derivatives of the joint survival function of lifetimes and censoring.

Example 1 Denote by (T, Δ), $T = \min(X, Y)$, $\Delta = 1\ (X \leq Y)$ censoring observations under the assumption of dependent censoring. Let $S(x, y)$ and $F(x, y)$ be the joint survival and probability density function of X and Y, respectively.

Consequently, the sub-distribution functions can be derived as follows:

$$H_1(t) = P(T < t\ , \Delta = 1) = P(X < t\ , X \leq y)$$

$$= \int_{x<t,} \int_{x \leq y} f(x,y)\, dxdy = -\int_0^t S_x[x,x]dx$$

This implies that the sub-density of a non-censored $(\delta = 1)$ shared observation is a derivative of sub distribution function.

$$h_1(t) = H_1'(t) = -S_x(t,t) = -\frac{\partial S}{dx}(x,y)\Big|_{y=t}^{x=t}$$

Similar calculations yield the sub-distribution and sub-density functions in the case of a censored observation ($\delta = 0$):

$$H_0(t) = P(T < t\ , \Delta = 0) = P(Y < t\ , Y < X)$$

$$= \int_{y<t,\ y<x} \int f(x,y)\, dxdy$$

and

$$h_0(t) = H_0'(t) = -S_y(t,t) = -\frac{\partial S(x,y)}{\partial y}\Big|_{y\ =\ t}^{x\ =\ t}$$

The likelihood function is a composition of the sub-density functions.

$$L(t,\delta) = -\delta S_x(t,t) - (1-\delta)S_y(t,t)$$

3.4 Kaplan Meier estimate for the survival function.

Suppose that they are n individuals with observed survival times $t_1 < t_2 < \cdots < t_n$ some of these observations may be right censored and may be more than one individual with the same observed survival time. Let's suppose that there are other r death times amongst the individuals, where $r \leq n$.

The number of individuals who are alive just before time $t_{(j)}$, including those who are about to die at this time, will be denoted by n_j, for j=1,2,…, r and d_j will denote the number who die at this time. The time interval from $t_{(j)-\sigma}$ to $t_{(j)}$, where Δ is an infinitesimal time interval, then includes one death time.

Since there are n_j individuals who are alive just before $t_{(j)}$ and d_j death at $t_{(j)}$, the probability that an individual dies during the interval from $t_{(j)} - \Delta t_{(j)}$ is estimated by the d_j/n_j.

The corresponding estimated probability of survival through that interval is given by $(n_j - d_j)/n_j$

In the limit, as $\Delta \rightarrow 0$, $(n_j - d_j)/n_j$ becomes an estimate of the probability of surviving from $t_{(j)}$ to $t_{(j+1)}$

Let us assume that the deaths of individuals in the sample occur independently of one another, the estimated survivor function at any time in the k^{th} constructed time interval from $t_{(k)}$ to $t_{(k+1)}$, $k = 1,2,\ldots,r$, $t_{(r+1)}$ is defined to be ∞, where will be estimated probability of surviving beyond $t_{(k)}$.

This is the Kaplan-Meier estimate of survival function, which is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{3.7}$$

For $t_{(k)} \leq t < t_{k+1}, k = 1,2,\ldots,r$ with $\hat{S}(t) = 1$ for $t < t_{(1)}$ and where $t_{(r+1)}$ is taken to be ∞.

3.5 Standard error of the Kaplan-Meier estimate

Since the Kaplan-Meier estimate is the most important and widely used estimate of the survivor function, $S(t)$, the standard error will be derived in this section 4.4.1

The Kaplan-Meier estimate of the survivor function for any value of t in the interval from $t_{(k)}$ to $t_{(k+1)}$ can be written as

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p_j} \tag{3.8}$$

For $k = 1,2,\dots,r$ where $\hat{P}_j = \frac{n_j - d_j}{n_j}$ is the estimated probability that an individual survives through the time interval which begins at $t_{(j)}$, $j = 1,2,\dots,r$.

Taking logarithms, in equation (3.8), we have

$$log\hat{S}(t) = \sum_{j=1}^{k} log\, \hat{P}_j$$

and so the variance of $log\hat{S}(t)$ is given by

$$Var\{\log \hat{S}(t)\} = \sum_{j=1}^{k} Var(\hat{P}_j) \tag{3.9}$$

Clearly, the number of individuals who survive through the internal beginning at $t_{(j)}$ can be assumed to have binomial distribution with parameters n_j and p_j where p_j is the true probability of survival through that interval. The observed number who survive is $n_j - d_j$, and using the result of variance $b(n,p)$, is $np(1-p)$, and the variance then variance is $n_j - d_j$ is given by

$$Var(n_j - d_j) = n_j p_j (1 - p_j)$$

Since $\hat{p}_j = (n_j - d_j)/n_j$ then the variance of $\hat{p}_j$ is

$$Var(\hat{p}_j) = Var\frac{(n_j - d_j)}{n^2{}_j}$$

Hence, the variance of $\hat{p}_j$ may be estimated by $\frac{\hat{P}_j(1-\hat{p}_j)}{n_j}$ (3.10)

Now using Taylor series approximation to the variance of a function of a random variable X, we have,

$$Var[g(X)] \approx \left\{\frac{d}{dx}g(X)\right\}^2 Var(X) \quad (3.11)$$

Using equation (3.11), the approximate estimated variance of $log\hat{P}_j$ is $\frac{(1-\hat{P}_j}{(n_j\hat{P}_j)}$, which on substitution for $\hat{P}_j$ (3.10), it reduces to

$$\frac{d_j}{n_j(n_j - d_j)}, \tag{3.12}$$

and from equation (3.9),

$$Var(log\hat{p}) \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

and a further application of the result in equation (3.11) gives

$$Var\{log\hat{S}(t)\} \approx \frac{1}{\left[\hat{S}(t)\right]^2} Var\left[\hat{S}(t)\right],$$

So that

$$Var\{\hat{S}(t)\} \approx \left[\hat{S}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_{j-}d_j)} \tag{3.13}$$

Finally, the standard error (s.e) of the Kaplan-Meier estimate of the survivor function, defined to be the square root of the estimated variance of the estimate, is given by

$$s.e.\ \{\hat{S}(t)\} \approx [\hat{S}(t)]\left\{\sum_{j=1}^{k}\frac{d_j}{n_j(n_j - d_j)}\right\}^{\frac{1}{2}} \tag{3.14}$$

for $t_{(k)} < t < t_{(k+1)}$.

Equation (3.11) is known Greenwood's formula.

If there are no censored survival times, and expression (3.12) becomes

$\frac{n_j - n_{j+1}}{n_j n_{j+1}}$. Now

$$\sum_{j=1}^{k}\frac{n_j - n_{j+1}}{n_j n_{j+1}}$$

$$= \sum_{j=1}^{k}\left(\frac{1}{n_{j+1}} - \frac{1}{n_j}\right)$$

$$= \frac{n_1 - n_{k+1}}{n_1 n_{k+1}}$$

Which can be written as

$\frac{1-\hat{S}(t)}{n_1\hat{S}(t)}$ since $\hat{S}(t) = \frac{n_{k+1}}{n_1}$

For $t_{(k)} \leq t < t_{(k+1)}, k = 1,2, \dots, r-1$, in the absence of censoring. Hence, from (3.13), The estimated variance of $\hat{S}(t)$ is $\frac{\hat{S}(t)[1-\hat{S}(t)]}{n_1}$ on assumption that the number of individuals at risk at time t has a binomial distribution with parameters $n_1,\ S(t)$.

3.6 Confidence interval for the values of survivor function

The confidence interval for the true value of the survivor function at time t is obtained by assuming that the estimated value of the survivor function at t is normally distributed with mean $S(t)$ and the estimated variance given by

$$Var\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_i - d_j)} \text{ in equation (3.13)}$$

From the asymptotic normality of $\hat{S}(t)$, it follows immediately that $100(1-\alpha)\%$ confidence interval for $S(t)$ (t fixed) is given by

$$\hat{S}(t) \pm z_{\frac{\alpha}{2}} \sqrt{Var(\hat{S}(t))} \qquad 3.15$$

The main drawback of this interval in equation (3.15) is the fact that it contains points outside the interval $[0,1]$

Hence, an appropriate transformation is used to determine the next confidence interval on the transformational scale and then back-transform.

One such appropriate transformation is the complementary log-log transformation $log(-logS(t)$. With $log(-logS(t))$ taking values between -∞ to ∞, the backward transformed value $S(t)$ takes the values between 0 and 1, as required.

The variance of $log(-logS(t)$ is obtained by applying the Taylor Series approximation of a function of a random variable in equation (3.11) leading to:

$$Var\left[\log(-log\hat{S}(t)\right] \approx \frac{1}{\left(log\hat{S}(t)\right)^2}\mathrm{Var}\left(log\hat{S}(t)\right)$$

$$\approx \frac{1}{\left(log\hat{S}(t)\right)^2}\sum_{j=1}^{k}\frac{d_j}{n_j(n_i - d_j)}$$

Back transforming the confidence interval for $log(-log\hat{S}(t))$ is:

$$\log\left(-log\hat{S}(t)\right) \pm z_{\frac{\alpha}{2}}\sqrt{Var\left[\log\left(-log\hat{S}(t)\right)\right]} \qquad (3.16)$$

leads to the following $100(1-\alpha)\%$ confidence interval for $S(t)$

$$\hat{S}(t)^{exp\left[\pm z_{\frac{\alpha}{2}}\sqrt{Var\left(log\left(-\log(\hat{S}(t))\right)\right)}\right]}$$

3.7 Estimating the survival function for left, truncated and right-censored data

For survival data where the event times of some subjects are also left-truncated, we need to define $R(y_i)$ as the number of individuals at risk at time $y_i$and under

observation prior to time y_i. Thus subject i is in risk set at event y_i. If $l_i \leq y_i < y_i$ where l_i and y_i are respectively the truncation time and event/censored time of subject i.

Because of the left-truncation, we cannot estimate $S(t)$ but only a conditional survival function of the form

$$S_l(t) = \frac{S(t)}{S(L=l)} = P(T > t | T > l)$$

as for the value of L. The first value we can possibly take is the lowest entry time in the study L; before that time the set is empty.

The survival distribution $S_l(t)$ is estimated by

$$\hat{S}(t) = \begin{cases} 1 & f \quad t < l \\ \prod_{j:l \leq y_i \leq t} \left(1 - \frac{d_i}{R(y_i)}\right) & if \ t \geq l \end{cases}$$

This estimator was first contributed by astronomer Lynden-Bell (1971) and can be derived by using counting process in the same way as for the censored data. It is possible for the number $R(y_i)$ to be quite small early in time. If for small y_i, all subjects in the risk set have an event, then $\hat{S}_L(t)$ will be zero for all t larger than y. This is not meaningful.

To circumvent this problem, one estimate the survival function condition on survival beyond a certain time point $l > t_1$, so that the risk set at that starting point is much larger. They observed events before that time will be neglected.

3.8 The Nelson Alien estimator for H(t).

Nelson (Nelson,1972) proposed an estimator for cumulative hazard $H(t)$ which was independently discovered by Aalen (Aan,1978) from the counting process approach. The Nelson-Aalen estimator is given by

$$\widehat{H}(t) = \sum_{y_i<t} \frac{d_j}{R(y_i)} \; for \; t \leq y_{(r)} \tag{3.17}$$

and its variance is given by

$$Var\left(\widehat{H}(t)\right) = \sum_{j:y_{(j)}\leq t} \frac{d_j}{R(y_i)^2} \tag{3.18}$$

We further have the following asymptotic result

$\frac{\hat{H}(t)-H(t)}{\sqrt{\hat{V}(H(t)}}$ is distributed $N(0,1)$

The survival estimator can be obtained from the Nelson-Aalen estimator from the relation $H(t) = -logS(t))$. This estimator is better than Kaplan-Meier estimator for small samples.

3.9 The median and percentile of the survival function

Instead of using the mean and as central measure, one often used the median in the context of survival analysis. The main reason is because the survival distribution is often skewed to the right, in which case a robust measure, like the median, is more appropriate. Moreover, the estimator of the mean is inconsistent, while the median can be estimated in a consistent way.

An estimator of the p^{th} quantile is given by

$$\bar{x}_p = inf\{t|\hat{S}(t) \leq 1 - p\}$$

an estimate of the median is given by $\bar{x}_p = 0.5$

It can be shown that

$$Var(\bar{x}_p) = \frac{(1-p)^2 Var(S(\bar{x}_p)}{S(\bar{x}_p)\hat{f}^2(\bar{x}_p)}$$

Where $\hat{f}$ is an estimator of the density f. Since the estimation of f involves smoothing techniques and the choice of a bandwidth sequence, we prefer not to use this variance estimator in the construction of a confidence interval for x_p but rather use a simple formula which is constructed as follows

Due to the asymptotic normality of $\hat{S}(x_p)$ it holds that

$$P\left(z_{\frac{\alpha}{2}}\right) \leq \left(\frac{\hat{S}(x_p) - S(x_p)}{\widehat{Var}(\hat{S}(t)} \leq -z_{\frac{\alpha}{2}}\right) = 1 - \alpha \qquad (3.19)$$

With $S(x_p) = 1 - p$. Therefore, a 95% confidence interval for x_p consist of all values t for which $\hat{S}(t)$ falls in this $100(1-\alpha)\%$ confidence region denoted in the equation (3.19). i.e., t satisfying the following condition

$$z_{\frac{\alpha}{2}} \leq \frac{\hat{S}(t) - S(x_p)}{\sqrt{Var(\hat{S}(t)}} \leq -z_{\frac{\alpha}{2}}$$

3.10 Hypotheses

3.10.1 Hypothesis for hazard function of one population.

We test whether a censored sample of size n comes from a population with a specified hazard function $h_0(t)$, with $h(t)$, the true hazard function for population under consideration, we wish to test.

$H_0 : h(t) = h_0(t)$ for all $t \leq y_i$,

$H_1 : h(t) \neq h_0(t)$ for some $t \leq y_i$,

Let $R(y_i)$ be the number of individuals at risk just before (prior) to event time y_i and d_j be the number of events at the ordered event times y_j, $j = 1,2, \ldots, k$.. The quantity $\frac{d_j}{R(y_i)}$ is crude estimator of the hazard function at the event time y_i (based on the Nelson-Aalen estimator of the cumulative function).

Under the null hypothesis, the hazard function at time y_j, is $h_0(y_j)$.

Let $w(t)$be some weight function with $w(t) = 0$ for $t > y_k$. The test statistics is given by

$$Z = \sum_{j=1}^{k} w(y_j) \frac{d_j}{R(y_1)} - \int_0^{y_k} w(s) h_0(s) ds$$

Under null hypothesis (H_0), the variance of statistics Z is given by

$$Var(Z) = \int_0^{y_k} w^2(s) \frac{h_0(s)}{R(s)} ds$$

With $R(s)$ corresponding to the number of subjects in the risk set at time s

for large samples, the statistics

$$\frac{Z}{\sqrt{Var(Z)}} \sim N(0,1)$$

which can be used to test both one-sided and two-sided hypothesis.

The most common choice for weighted function is $w(t) = R(t)$ which leads to the test.

$$Z = \sum_{j=1}^{k} d_j - \int_0^{y_k} R(s) h_0(s) ds \qquad (3.20)$$

The second term of the right-hand-side of equation (3.20) can be rewritten as a sum over the different observations leading to:

$$Z = \sum_{j=1}^{k} d_j - \sum_{i=1}^{n} \int_0^{y_i} h_0(s)ds$$

$$= \sum_{j=1}^{k} d_j - \sum_{i=1}^{n} H_0(y_i)$$

$$= O -$$

E (3.21)

Which corresponds to the one-simple log-rank test

Clearly, O in equation 3.21 stands for the number of observed events. i.e.,

$$O = \sum_{j=1}^{k} d_j$$

On the other hand E is related to the expected number of events under the null hypothesis. For the one-simple log-rank statistic with $w(t) = R(t)$, the variance is given by

$$Var(Z) = \int_0^{y_i} R(s)h_0(s)ds = E$$

and thus we have

$$\frac{O - E}{\sqrt{E}} \sim N(0,1)$$

A less common used weight function

$$w(t) = R(s)S_0^p(t)\ (1 - S_0(t)^2 \qquad p, q \geq 0$$

has been proposed by Harrington and Fleming (1982).

Remark: More weight can be put on early deviations from H_0 by taking $p >> q$, on late deviations from H_0 by taking $p << q$ and deviation on the middle by taking $p = q > 0$

With the choice $p = q = 0$ we obtain the special case of the log-rank test, and $p = 1,\ q = 0$ the statistics is a generalization of the one-sample Wilcoxon test to the censored data.

3.10.2 Comparing the hazard functions of two or more populations.

We will first consider the comparison of the hazard function $h_1(t)$ and $h_2(t)$ of two populations

Testing the hypothesis.

$H_0 : h_1(t) = h_2(t)$ for all $t \leq y_k$,
$H_1 :\ h_1(t) \neq h_2(t)$ for some $t \leq y_k$, (3.22)

Let $y_1 < y_2 < \cdots < y_k$ be the event times in the pooled sample,

Where $d_k(k)\ (k = 1,2)$, the number of events at the time y_i in sample K and $R\left((y_j(k)\right)$ the number of individuals at risk at time in the sample k.

Further let $d_j = \sum_{k=1}^{2} d_j K \ \ and \ R(y_j) = \sum_{k=1}^{2} R\left(y_j(k)\right)$ be the number of events and the number of risk in the pooled sample at time y_j.

For each event time a 2×2 contingency table as displayed in Table 3.1 can be derived.

Table 3. 1 : The 2×2 contingency table for the event time

Group	Event	No Event	Total
1	$d_j(1)$	$R(y_j(1)) - d_j(1)$	$R(y_j(1))$
2	$d_j(2)$	$R(y_j(2)) - d_j(2)$	$R(y_j(2))$
Total	d_j	$R(y_j) - d_j$	$R(y_j)$

For the testing H_0, the quality of the hazard function in the two populations, we need the independence between rows and columns, which corresponds to the assumption that the hazard in the two groups at time y_i is the same. A typical test statistics used for such 2×2 contingency table at time y_j, choosing Group1 as reference group, is given by.

$$O_j - E_j = d_j(1) - \frac{d_j R\left(y_j(1)\right)}{R(y_i)} \tag{3.23}$$

Where O_j corresponds to the observed number of events in the first group (group1) and E_j to the expected number of events in the first group, given that the hazard is equal in the two groups.

The test statistics U is defined as a weighted average of the terms (3.23) over the different event times

$$U = \sum_{j=1}^{k} w(y_i)\big(O_j - E_j\big) = \sum_{j=1}^{k} w(y_i)\left(d_j(1) - \frac{d_j R\big(y_j\big)}{R\big(y_j\big)}\right) \tag{3.24}$$

The variance of the test statistics can be obtained by observing that conditional on $d_j, R(y_j)$ and $R(y_j)$, the statistics $d_j(1)$ has hypergeometric distribution.

3.11 Estimating the cumulative Hazard function

The hazard function is more informative about the underlying population than the survival or cumulative hazard function (for two populations for which survival functions have about the same shape, the corresponding hazard functions might be completely different).

A crude way (of obtaining) to obtain an estimate of the hazard function is to take the sizes of the jumps of the cumulative hazard as the estimate $\hat{h}(t)$. This corresponds then to the discrete function with only a value greater than zero at event time.

A better estimator can be obtained by smoothing-techniques. Suppose we want to estimate the hazard at time t. Since time points close to the point t are informative, The smooth estimator of $h(t)$ takes a (weighted) average of the crude estimates at time points lying in the interval for $[t-b, t+b]$ for a certain value b, called the bandwidth.

One such example is

$$h(t) = \frac{1}{(2b)} \sum_{j=1}^{k} I\big(-b \le t - y_j \le b\big)\Delta\widehat{H}\,(y_i) \tag{3.25}$$

Where $\widehat{H}(t)$ is any estimator of the cumulative hazard rate and $\Delta\widehat{H}(y_i)$ is the estimated increase of the cumulative hazard in y_i

More generally, we define

$$\hat{h}(t) = \frac{1}{b}\sum_{j=1}^{k} K\left(\frac{t - y_j}{b}\right)\Delta\widehat{H}\,(y_i) \tag{3.26}$$

where $K(.)$ is a Kernel density function.

Some examples of Kernel density are given in Table 3.2.

Table 3. 2: Kernels for smoothing functions of hazard estimation.

Name	Density function	support
Uniform	$K(x) = \frac{1}{2}$	$-1 \leq X \leq 1$
Epanechnikov	$K(x) = \frac{3}{4}(1 - x^2)$	$-1 \leq X \leq 1$
$b_i\ weight$	$K(x) = \frac{15}{16}(1 - x^2)^2$	$-1 \leq X \leq 1$

The variance of $\hat{h}(t)$ is given by

$$Var\left(\hat{h}(t)\right) = \frac{1}{b^2}\sum_{j=1}^{r} K\left(\frac{t - y_j}{b}\right)^2 \Delta\widehat{Var}\left(\widehat{H}(y_i)\right) \tag{3.27}$$

Where

$$\Delta\widehat{Var}\left(\widehat{H}(y_i)\right) - Var\left(\widehat{H}(y_{j-1})\right)$$

Remark: • The uniform Kernel corresponds with Kernel given in equation (3.26).

• The last two Kernels in table 3.2 assign more weight to observations close to the considered time point.

Remark. The hypergeometric distribution is the counterpart of the binomial distribution for sampling in a finite population without replacement.

The probability of X is then given by the hypergeometric distribution.

$$P(X = x|N, m, n) = \frac{\binom{m}{x}\binom{N-m}{n-k}}{\binom{N}{n}}$$

$$For\ max(0, n - m - N) < X < min(m, n)$$

Therefore, we have $E(x) = \frac{mn}{N}$ and

$$Var(X) = \frac{n(\frac{m}{n})(1 - \frac{m}{N})(N - n)}{N - 1}$$

Now applying the variance of the hypergeometric distribution in equation (3.27) to $d_j(1)$ in 2×2 in contingency, Table 3.2, we obtain.

$$Var\left(d_j(1)\right) = \frac{d_j\left(\frac{R(y_j(1)}{R(y_j)}\right)\left(1 - \frac{R(y_j(1)}{R(y_j)}\right)\left(R(y_j) - d_j\right)}{R(y_i) - 1} \quad (3.28)$$

Table 3. 3: The 2×2 contingency table for hypergeometric distribution

	Event	Sampled	Not sampled
Defective	x	$m-x$	m
Non-defective	$n-x$	$N+x-n-m$	$N-m$
Total	n	$N-n$	N

And thus

$$Var(U)=\sum_{j=1}^{k} w^2(y_i)Var(d_j(1) \tag{3.29}$$

As we have conditioned on $d_j, R\left(y_j(1)\right)$ and $R(y_j)$

It follows that for large samples and under null hypothesis

In a practical data analysis, the choice of weight is extremely important, as have been proposed in literature. The most common weight functions are:

$$\frac{U}{\sqrt{Var(U)}} \sim N(0,1)$$

$w(y_j)=1$ this choice of weight function leads to log-rank test and has optimum power to detect alternatives when the hazard rates in the different populations are proportional to each other.

$w(y_i)=R(y_j)$ this weight function yields the generalization of Gahan (1965) of two sample. Mann-Whitney-Wilcoxon and generalization of Breslow (1970) of the Kruskal-Walis test. The weight put more emphasis on time points with many data. It

further depends on censoring distribution. However, this type of weight can behave badly in some circumstances.

$w(y_j) = f(R((y_j))$(with f a fixed function). This weight is proposed by Tarone and Ware (1977). A suggested choice for the function is $f(R\left((y_j)\right) = \sqrt{R(y_j)}$. This class of weight gives more weight to differences with still many subjects at risk. This weight also depends heavily on the event times and censoring distribution.

$w(y_j) = \hat{S}(y_j)\frac{R(y_k)}{R(y_k)+1}$. This weight function was proposed by Peto and Peto (1972) and Kalbfleisch and Prentice (1980) is modified slightly as above. This weight function depends on the combined survival experience in the pooled sample and these weights are predictable.

$w(y_j) = \left(\hat{S}(y_j - 1)\right)^p \left(\hat{S}(y_j - 1)\right)^q, p \geq 0, q \geq 0$ The weight function is Kaplan-Meier estimator of the pooled sample. This survival function at previous event time is used as a weight so that they are known prior to the time at which the comparison is made. Fleming and Harrington (1981) proposed this general class of weights that include, as special cases, the weight used in the log-rank test in the version of the Mann-Whiteney-Wilcoxon test $(p = 1, q = 0)$ very close to the suggested by Peto and Peto (1972) with $q = 0, p > 0$, more weight is put on early differences between the hazard functions, with $p = 0\ and\ q > 0$, more weight is put on the late differences.

Table 3. 4: Weighting Schemes for Different Survival Analysis Tests

Test	
Log-rank	$w(y_j) = 1$
Wilcoxon	$w(y_j) = R(y_j)$
Peto-Peto	$w(y_j) = \hat{S}(y_j)\frac{R(y_k)}{R(y_j)+1}$
Harrington-Flemming	$w(y_j) = [\hat{S}(y_j - 1)]^p[1 - \hat{S}(y_{j-1})]^q$
Tarone-Ware	$w(y_j) = f(R((y_j))$

Remark: An important decision in comparing survival function is the choice of weight function. The choice of weight function should never be based on the observed data. The decision should be made before looking at the data.

3.12 Comparing more than two populations.

We are now interested in testing hypotheses

$$H_0 : h_1(t) = h_2(t) = \cdots = h_l(t) \; for \; all \; t \leq y_k$$

$$H_1 : h_i(t) \neq h_j(t) \; for \; at \; least \; one \; pair \; (i,j) for \; some \; t \leq y_k \qquad (3.30)$$

Where $h_1(t), h_2(t), \ldots, h_l(t)$ are hazard functions of populations $k = 1,2,\ldots,l$. Let $y_1 < y_2 < \cdots < y_k$ be ordered event times in the pooled sample, $d_j(k)$ $(j = 1,2,\ldots,r \; and \; k = 1,2,\ldots,l)$. The number of events at the time y_j in sample k, and $R(y_j(k)$ the number of individuals at risk y_j in the sample of population k.

Further, let

$$d_j = \sum_{k=1}^{l} d_j(k) \quad and\ R(y_j) = \sum_{k=1}^{p} R\big(y_j(k)\big)$$

be the number of events and the number of risk, respectively, in the pooled sample at time y_k.

Hence, the test statistics is now based on the $l \times 2$ contingency tables for different event times y_j as in table 3.4.

Table 3. 5: The l×2contingency table for event time containing the number of subjects with and without event at time in the l groups and the number at risk just before that time.

Group	Event	No event	Total
1	$d_j(1)$	$R\big((y_j(1)\big) - d_j(1)$	$R\big((y_j(1)\big)$
2	$d_j(2)$	$R\big((y_j(2)\big) - d_j(2)$	$R\big((y_j(2)\big)$
3	$d_j(3)$	$R\big((y_j(3)\big) - d_j(3)$	$R\big((y_j(3)\big)$
. . .	. . .	. . .	. . .
l	$d_j(l)$	$R\big((y_j(l)\big) - d_j(l)$	$R\big((y_j(l)\big)$
Total	d_j	$R(y_j) - d_j$	$R(y_j)$

Let random vector $D_j^t = \big(D_j(1), D_j(2), \dots, D_j(l)\big)$ has a multivariate hypergeometric distribution.

$$P(D_j(1) = d_j(1), \dots, d_j(l) = d_j(L) = \frac{\prod_{k=1}^{l} \binom{R((y_j(k))}{d_j(k)}}{\substack{R(y_j) \\ d_j}}$$

with

$$E(D_j) = E\left(D_j(1), D_j(2), \dots, D_j(l)\right)$$

$$= \left(R(y_j(1)\frac{d_j}{R(y_j)}, \dots, R(y_j(l)\frac{d_j}{R(y_j)}\right)$$

and the variance of D_j contains the variance on the diagonal and covariance off-diagonal. The entries of the $Var(D_j)$ are given by

$$Cov\big((D_j(k), D_j k')\big) = \frac{R(y_j(k)d_j\big(R(y_j) - d_j\big)}{R\big(y_j\big)R\big(y_j\big) - 1}\left(I(k = k') - \frac{R(y_j k')}{R(y_j)}\right) \quad 3.31$$

For population k, we define

$$U_k = \sum_{j=1}^{r} w(y_j)\left(d_j k - \frac{d_j R(y_j(k)}{R(y_j)}\right) \quad 3.32$$

which is the weighted sum of the differences between the observed and the expected number of events under H_0.

The components of the rector $(u_1, u_2, \dots, u_l)$ are linearly dependent because

$$\sum_{k=1}^{l} U_k = 0$$

Indeed

$$\sum_{k=1}^{l} U_k = \sum_{k=1}^{l}\left[\sum_{j=1}^{r} w(y_j)\left(d_j k - \frac{d_j R\left(y_j(k)\right)}{R(y_j)}\right)\right]$$

$$\sum_{j=1}^{r} w(y_j)\left[\sum_{k=1}^{l}\left(d_j k - \frac{d_j R\left(y_j(k)\right)}{R(y_j)}\right)\right]$$

$$\sum_{j=1}^{r} w(y_j)\left(dy_j - \frac{d_j R(y_j)}{R(y_j)}\right)$$

$$= 0$$

Therefore we select the first $l-1$ components leading to the vector $U^t = (u_1, \ldots, u_{l-1})$

Furthermore, using equation (3.29), it follows that the entries of the variance-covariance matrix u, are given by

$$Cov(u_k, u_k') = \sum_{j=1}^{r} w^2(y_j)\, dy_j - \frac{R\left(y_j(k)\right) d_j\left(R(y_j) - d_j\right)}{R(y_j)\left(R(y_j) - 1\right)}$$
$$- \left(I(k = k') - \frac{R\left(y_j k'\right)}{R\left(y_j\right)}\right) \quad (3.33)$$

The test statistic $U^t V^{-1}(U)$ has a Chi-square distribution with $l-1$ degrees of freedom under the null hypothesis and for large sample sizes.

3.13 Test for trend.

The hypothesis to test for trend is defined by

$$H_0 : h_1(t) = h_2(t) = \cdots = h_l(t) \; for \; all \; t \leq y_k$$

$$H_1 : h_1(t) \leq h_2 \leq (t) < h_l(t) \; for \; some \; t \leq y_k \; with \; at \; least \; one \; strict \; inequality \quad (3.34)$$

The alternative hypothesis in equation (3.30) is equivalent to the hypothesis $S_1(t) \geq S_2(t) \ldots > S_l(t)$

an appropriate test statistic for the trend is given by

$$U = \sum_{k=1}^{l} w_k u_k \; with \; u_k$$

in Summary statistics in equation (3.32) for the k^{th} population and w_k the code assigned to the k^{th}population. The population codes, $w_1, w_2, \ldots, w_l$ are often considered equal spaced to correspond to a linear trend across the groups.

The variance of u is given by

$$Var(U) = \sum_{k=1}^{l} \sum_{k'=1}^{l} w_k w_k' \; Cov\,(u_k, u_k')$$

using equation (3.28) for $Cov(u_k, u_k')$.

with sufficiently large sample size, testing can be based on $\frac{U}{\sqrt{Var(U)}} \sim N(0,1)$

3.14 Stratified tests

It often happens in practice that the subjects in a study are grouped according to their prognosis, resulting, for instance, in good, average and poor diagnosis groups, such groups are called strata in the context of survival analysis, and are similar to block factors are encountered in linear model. It is often advisable to adjust for the strata, as it might reduce variance considerably, thereby increasing the power of the test.

Define u_{kb}, $k = 1,2,\ldots,l$, $b = 1,2,\ldots,m$ as a summary statistic for population k obtained as in equation (3.32) but now referring to stratum b.
The stratum summary statistics for population k is then given by

$$U_k = \sum_{b=1}^{m} U_{kb}$$

We collect the first $l-1$ stratified summary statistics in the vector $U^t = (u_1, \ldots, u_{l-1})$

The entries of the variance-covariance matrix of u, $Var(u)$ are given by

$$Cov(u_k u_{k'}) = \sum_{b=1}^{m} Cov(u_{kb}, u_{k'b})$$

With $Cov(u_{kb}, u_{k'b})$ obtained as in equation (3.32) but now referring to stratum b.

The test statistics $U^t V^{-1}(U)U$ now has a Chi-square distribution with $l-1$ degrees of freedom under null hypothesis and for large sample sizes.

If the comparison is based on two populations, testing is based on the following distribution property.

Where U_b and $Var(U_b)$ are obtained as in equation (3.23) and (3.26) respectively, but now within stratum b.

3.15 Modelling survival data

Through a modelling approach to the analysis of survival data, researchers can explore how the survival experience of a group of patients depends on the value of one or more explanatory variables, where values were recorded for each patient at the time of origin. The aim is to determine which of the explanatory variables have an impact on the survival time of the patients. Another aim is to identify whether patients in the two treatment groups have different survival experiences. In this study, additional variables such as the age of the patient and the size of their tumour are likely to influence survival times, it will be important to consider these variables when assuming the extent of any treatment difference.

In the analysis of survival data, interest centres on the risk or hazard of death at any time after the origin of the study. As a consequence, in survival analysis, the hazard function is modelled directly. The resulting models are somewhat different in the form from linear models encountered in regression analysis and in the analysis of data

from designated experiments, where the dependence of the mean response, or some function of it, on certain explanatory variables it's modelled.

There are two broad reasons for modelling survival data. One objective of modelling process is to determine which combination of the potential explanatory variables affect the form of hazard function. In particular, the effect that the treatment has on the hazard is studied, as can the extent to which other explanatory variables affect the hazard function.
Another reason for modelling hazard function is to obtain an estimate of the hazard function itself for individual. In this study, the quality quantity such as median survival time, which will be a function of the explanatory variables in the model. The median survival time could be estimated for current or future patients with particular values of the explanatory variables. The resulting estimate is useful in devising a treatment region or counselling patients about their prognosis.

The basic model for survival data that is considered in this study is semi-parametric model. This model was proposed by Cox (1972)

3.16 The semi-parametric proportional model

The semi-parametric proportional model is defined as

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}^t \boldsymbol{x}_i) \tag{3.35}$$

With $\boldsymbol{x}_i$ the $\rho x1$ a vector containing the covariate information for subject i, the parameter vector and $h_0(t)$ the baseline function, which corresponds to the hazard function for a subject with $\boldsymbol{x}_i = \mathbf{0}$. In Cox model, the baseline, $h_0(t)$, is unspecified nuisance function that makes the model same-parametric.

We assume that the ratio of two subject with covariate information $\boldsymbol{x}_i$and $\boldsymbol{x}_j$ is constant over time.

i.e.,

$$\frac{exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)}{exp(\boldsymbol{\beta}^t \boldsymbol{x}_j)} = \Psi$$

3.17 Partial likelihood for data without ties

Let, with r the number of observed event times (assuming no ties and thus $= d$, the number of events),$y_1 < y_2 < \cdots < y_r$ denote the orded event times with corresponding covariate $\boldsymbol{x_1} < \boldsymbol{x_2} < \cdots < \boldsymbol{x_r}$. The survival likelihood can then be written as

$$\prod_{j=1}^{r} h_0(j) \exp\left(\boldsymbol{x}_{(j)}^t \boldsymbol{\beta}\right) \prod_{i=1}^{n} \exp\left(-H_0(y_i)\right) \exp\left(\boldsymbol{x}_{(j)}^t \boldsymbol{\beta}\right) \tag{3.36}$$

With $h_{0j} = h_0(y_j)$

Based on non-parametric maximum likelihood estimation ideas for right-censored data, it is natural to work with the following discrete version of cumulative baseline hazard:

$$H_0^{disx}(y_i) = \sum_{y_i < y_j} h_0\left(y_j\right) \tag{3.37}$$

This implies that we take $h_0(t) = 0$ except for times at which an event occurs, as this choice leads to the largest contribution to the likelihood if a discrete hazard function is assumed.

Using equation (3.36) with $\boldsymbol{\beta}$ fixed we obtain, after rearranging terms, that the survival likelihood in equation (3.35) can be written as

$$L(h_0(1), h_0(2), \dots, h_0(r)|\boldsymbol{\beta}) = \prod_{\boldsymbol{j=1}}^{\boldsymbol{r}} \big(h_{\boldsymbol{0}}(j)\big) \prod_{j=1}^{r} \exp\big(\boldsymbol{x}_{(j)}^{t}\boldsymbol{\beta}\big) \prod_{j=1}^{r} \exp\big(-h_{\boldsymbol{0}}(j)\big) \sum_{k \in R(y_{(j)})} \exp\big(\boldsymbol{x}_{(j)}^{t}\boldsymbol{\beta}\big) \quad (3.38)$$

Where $R\big(y_{(j)}\big)$is the risk set a time $y_{(j)}$ containing all subjects that are still at risk to experience the event at that time.

Taking partial derivative with respect to $h_0(1)$ we obtain

$$\frac{\partial L(h_0(1), h_0(2), \dots, h_0(r)|\boldsymbol{\beta})}{\partial h_0(1)}$$

$$= \prod_{j=1}^{r} \exp\big(\boldsymbol{x}_{(j)}^{t}\boldsymbol{\beta}\big) \prod_{j=1}^{r} \exp\big(-h_0(j)b_j\big)(h_0(2), \dots, h_0(r) - h_0(1)h_0(2), \dots, h_0(r), b_1)$$

with

$$b_j = \sum_{k \in R(y_{(j)})} \exp(\boldsymbol{x}_k^{t}\boldsymbol{\beta})$$

Equating this partial derivative to zero we have

$$1 - h_0(1) = 0$$

solving for $h_0(j)$ gives

$$h_0(j) = \frac{1}{b_j} = \frac{1}{\sum_{k \in R(y_{(j)})} \exp(x_k^t \boldsymbol{\beta})} \tag{3.39}$$

Plugging equation (3.39) into equation (3.38)

we obtain upon a factor $\exp(\alpha)$ which does not contain any of the parameters,

$$L(\boldsymbol{\beta}) = \prod_{\boldsymbol{j=1}}^{\boldsymbol{r}} \frac{\exp(x_k^t \boldsymbol{\beta})}{\sum_{k \in R(y_{(j)})} \exp(x_k^t \boldsymbol{\beta})}$$

$L(\boldsymbol{\beta})$ is called partial likelihood

This expression is used to estimate $\boldsymbol{\beta}$ through maximisation

Using the logarithm of the partial likelihood

$$L(\boldsymbol{\beta}) = \sum_{k \in R(y_{(j)})}^{r} \exp(\boldsymbol{x}_k^t \boldsymbol{\beta}) - \sum_{j=1}^{r} log \left(\sum_{k \in R(y_{(j)})} \exp(x_k^t \boldsymbol{\beta}) \right) \tag{3.40}$$

Parameter estimate can then be obtained by maximising equation (3.40)

To maximise (3.40) we will use Newton-Raphson procedure, which is based on the following iterative procedure

$$\widehat{\boldsymbol{\beta}}_{new} = \widehat{\boldsymbol{\beta}}_{old} + I^{-1}(\widehat{\boldsymbol{\beta}}_{old}) \boldsymbol{U}(\widehat{\boldsymbol{\beta}}_{old})$$

With $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_{old})$ and $I^{-1}(\widehat{\boldsymbol{\beta}}_{old})$ the vector scores and the inverse of the observed information matrix respectively. Convergence is reached when $\widehat{\boldsymbol{\beta}}_{old}$ and $\widehat{\boldsymbol{\beta}}_{new}$ are

sufficiently close together. In order to use Newton-Raphson procedure we will require $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_{old})$ and $I^{-1}(\widehat{\boldsymbol{\beta}}_{old})$

The score function for parameter β_h is given by

$$U_{\boldsymbol{h}}(\boldsymbol{\beta}) = \frac{\boldsymbol{\delta} l(\boldsymbol{\beta})}{\boldsymbol{\delta}\beta_h} = \sum_{j=1}^{r} x_{(j)}h - \sum_{j=1}^{r} \frac{\sum_{k\in R(y_{(j)})} x_{kh} \exp(x_k^t \boldsymbol{\beta})}{\sum_{k\in R(y_{(j)})} \exp(x_k^t \boldsymbol{\beta})} \tag{3.41}$$

The observed information matrix is the negative of the matrix of second derivatives of the log likelihood and is given by $\boldsymbol{I}(\boldsymbol{\beta})$ with the entry in column h and row l given by

$$\frac{\delta^2 l(\boldsymbol{\beta})}{\delta\beta_h\beta_l} = \sum_{j=1}^{r} \sum_{k\in R(y_{(j)})} x_{kh}x_{kl} \exp(\boldsymbol{x}_k^t \boldsymbol{\beta}) - \sum_{j=1}^{r} \frac{\sum_{k\in R(y_{(j)})} x_{kh} \exp(x_k^t \boldsymbol{\beta})}{\sum_{k\in R(v_{(j)})} \exp(x_k^t \boldsymbol{\beta})} \times \sum_{j=1}^{r} \left[\frac{\sum_{k\in R(y_{(j)})} x_{kl} \exp(x_k^t \boldsymbol{\beta})}{\sum_{k\in R(y_{(j)})} \exp(x_k^t \boldsymbol{\beta})} \right] \tag{3.42}$$

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be approximated by the inverse of the information matrix, evaluated at $\widehat{\boldsymbol{\beta}}$, i.e., $I^{-1}(\widehat{\boldsymbol{\beta}})$

The variance of $\widehat{\boldsymbol{\beta}}_h$, $Var(\widehat{\boldsymbol{\beta}}_h)$ can thus be approximated by the entry of matrix $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})$ on the diagonal column and row h

The properties (consistency, asymptotic normality) of the partial likelihood estimator for interval for $\widehat{\boldsymbol{\beta}}_h$ is given by

$$\widehat{\boldsymbol{\beta}}_h \pm Z_{\frac{\boldsymbol{\alpha}}{\boldsymbol{2}}} \sqrt{Var(\widehat{\boldsymbol{\beta}}_h)}$$

from which the $100(1-\alpha)\%$ confidence interval for the hazard ratio Ψ_h follows as

$$exp\left(\widehat{\boldsymbol{\beta}}_h \pm Z_{\frac{\boldsymbol{\alpha}}{\mathbf{2}}}\sqrt{Var(\widehat{\boldsymbol{\beta}}_h)}\right)$$

3.18 Partial likelihood for data with tie observations

The censored observations that are tied with event time can therefore be dealt with in the partial likelihood equation (3.40). The situation, however, is different for two or more subjects that have an event exactly at the same time. Kalbfleisch and Prentice (1980) derived the approximate likelihood function for tie observations.

Due to its complexity, however, this likelihood function is rarely used, but rather approximation, and, one such approximation,

$$\prod_{j=1}^{r}\frac{\prod_{l:=y_{(j)},}\delta_l = 1\exp(\boldsymbol{x}_l^t\boldsymbol{\beta})}{\left[\sum_{k\leq y_{(j)}}\exp(\boldsymbol{x}_l^t\boldsymbol{\beta})\right]^{d_{(j)}}} \tag{3.43}$$

is due to Breslow (1974). Each subject with event time $y_{(j)}$ is therefore considered to have an event time distinct from all the other subjects in the risk set, with the subjects experiencing also an event at, time $y_{(j)}$, the event times of these other subjects are considered to be after that subject. This was obviously incorrect.

Efron (1977) proposes another approximation

$$\prod_{j=1}^{r}\frac{\prod_{l=y_{(j)}}\exp(x_l'\boldsymbol{\beta})}{\left[\prod_{h=l}^{d_{(j)}}\left(\sum_{k\leq y_{(j)}}\exp(x_k^t\boldsymbol{\beta}) - \frac{h-1}{d_{(j)}}\sum_{l=y_{(j)},\delta=1}\exp(x_l^t\boldsymbol{\beta})\right)\right]} \tag{3.44}$$

In which the subjects that also have an event at time $y_{(j)}$ do not contribute fully to the denominator. This approximation is closer to the exact partial likelihood given by Kalbfleisch and Prentice (1980)

A third approximation, due to Cox (1972), is based on the logistic discrete version of proportional hazards model (3.35)

$$\frac{h_i(t)}{1-h_i(t)} = \frac{h_0(t)}{1-h_0(t)} \exp(\boldsymbol{x}_l^t \boldsymbol{\beta}) \tag{3.45}$$

Which tends to model (3.35) as the discrete time interval tends to zero. For this discrete model, the exact likelihood function is given by

$$\prod_{j=1}^{r} \frac{\prod_{l:y_{l=y_{(j)},\delta=1}} \exp(x_l^t \boldsymbol{\beta})}{\sum_{q \varepsilon Q_j} \sum_{h \varepsilon q} \exp(x_h^t \boldsymbol{\beta})} \tag{3.46}$$

Where Q_j is the set of all possible combinations of $d_{(j)}$ subjects from the risk set the set $R(y_{(j)}$. The set contains

$$\binom{R(y_{(j)})}{d_{(j)}} \text{ elements,}$$

3.19 Testing hypothesis

The hypotheses are tested within the framework of the semi-parametric model.

Let $\boldsymbol{\beta}$ the $\rho x1$ vector, the global null hypothesis is given by

H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}_0$ is a vector containing the hypothesized values of the population parameters under the null hypothesis. Under the alternative hypothesis, at least one of these population parameter differs from the hypothesised value.

With $\widehat{\boldsymbol{\beta}}$ the maximum likelihood estimator and $I(\widehat{\boldsymbol{\beta}})$ the observed information matrix, a first test statistic, U_w^2, the Wald test statistic has the following properties under H_0 and with large sample size

$$U_{\boldsymbol{w}}^{\mathbf{2}} = (\widehat{\boldsymbol{\beta}} - \beta_0)^t \mathrm{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \beta_0) \sim \chi_p^2 \tag{3.47}$$

The second test statistics, U_{LR}^2 is the likelihood ratio test statistic with the following asymptotic result under H_0

$$U_{LR}^2 = 2\big(l(\widehat{\boldsymbol{\beta}})\big) - l(\boldsymbol{\beta_0}) \sim \chi_p^2 \tag{3.48}$$

with $l(\widehat{\boldsymbol{\beta}})$, $l(\boldsymbol{\beta_0})$ the log likelihood evaluated at $\hat{\beta}$

A third is based on the score test statistic, U_{sc}^2, with the following asymptotic result under H_0

$$U_{sc}^2 = U(\beta_0^t) I^{-1}(\beta_0) U(\beta_0) \sim \chi_p^2 \tag{3.49}$$

with $U(\beta_0)$ the score vector evaluated at β_0

3.20 Modelling building strategies for the parametric models

In many practical applications, a whole set of covariates is available to be included in the model. We will typically start fitting all univariable models. That way we will be able to identify candidate covariates to be considered for the multivariable model. Covariates with significant effects at the 10% or even 20% significance level are considered further. We can then choose for forward or backward selection procedure.

We have made use of the likelihood ratio test in equation (3.46), as models are nested in these procedures. However, if we need to compare models that are not nested, we will do so by Akaike's Information Criteria (AIC).

$$AIC = 12\log(L) + kp$$

where p is the number of parameters in the model, L is the likelihood is a constant. An often chosen value is $k = 2$

3.21 Diagnostic for the semi-parametric model

We will use diagnostic technique for the semi-parametric model to check whether the underlying assumptions are fulfilled, as model assumptions are checked based on residuals, we will select residuals relevant to our studies.

3.22 Residuals for the semi-parametric model

We will determine residual for the following simple model

$$h_i(t) = h_0(t)\exp(\beta^t \boldsymbol{x}_i) \tag{3.50}$$

Most of the residuals are a function of an estimate of the cumulative baseline hazard $\widehat{H}_i(t)$. We will use the Breslow (1974)'s estimator and derive Cox-Snell residuals, the modified Cox-Snell residual and Schoenfeld residual for our studies.

$$r_{ci} = \exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)\,\widehat{H}_0(y_i)$$

$$= -\log\left(\hat{S}_i(t)\right) \tag{3.51}$$

Let us assume that no censoring occurs. We consider the distribution of the underlying cumulative function $H_i(t) = -\log\left(\hat{S}_i(t)\right)$. With T the random variable expressing time to event with density function $f(t)$ the density function of $H_i(T) = -\log\big(S(T)\big)$

is given by

$$f_{H(T)}(H(t)) = f_T\left(S^{-1}\big(\exp(-H(T))\big)\right) \times \left\|\frac{d(-logS(t)}{dt}\right\|^{-1}$$

$$= f_T\left(S^{-1}\big(\exp(-H(t))\big)\right)\frac{S(t)}{f_T(t)}$$

$$= f_T\left(S^{-1}\left(\exp(-H(t))\right)\right)\frac{S\left(S^{-1}\left(\exp(-H(t))\right)\right)}{f_T\left(S^{-1}\left(\exp(-H(T))\right)\right)}$$

$$= \exp(-H(t))$$

which correspond to an exponential distribution with mean one, whatever the form of $S(t)$

If the model is appropriate, the estimate $\hat{S}_i(t)$ will be close to the underlying survival function $S_i(t)$. We expect that the sample of values $r_{ci} = -\log\left(\hat{S}(t)\right)$ resembles a draw from a unit exponential distribution.

In the case of right-censored data, the Cox-Snell residuals are also right-censored, the residual larger than the value obtained by using the censoring time y_i in equation (3.46). The residuals in the right censored data therefore constitute a censored sample of the unit exponential distribution.

To accommodate for right censoring, (Cox and Snell, 1968) proposed the modified Cox-Snell residual. Cox and Snell proposed to add an excess residual term for the censored observations leading to

$$r_{mci} = rc_i + (1 - \delta_i)\Delta \qquad (3.52)$$

To determine an appropriate value Δ, the lack of memory property of the exponential distribution is used, since has a unit exponential distribution, so has Δ, from which follows that $E(\Delta) = 1$. They (Cox and Snell) to put Δ in equation (3.45) equals to 1.

However, Crowley and Hu (1977) found that Δ= 1 inflates the variance too much, and therefore proposed the median of Δ= $log2$.

This residual in equation (3.52) are not symmetrically distributed around zero and $\widehat{H}_i(t) \geq 0$.

3.23 Schoenfeld residuals

Schoenfeld residuals have been proposed by Schoenfeld (1982). For each subject there is row a set of p residuals (with p the number of parameters) instead of just one residual).

The Schoenfeld residual for parameter k and subject i is given by

$$r_{sik} = \delta_i \left(x_{ik} - \sum\nolimits_{l \epsilon R(y_i)} x_{lk} \exp(\hat{\beta}_{x_l}^t) \right) \tag{3.53}$$

The residual corresponds to the difference between the covariates of the subject with failure and the expected value of the covariate at that time.

The only uncensored observations have score residuals different from zero. The residuals $r_{s_{ik}}$ are termed partial residuals because they can be obtained from partial derivative of the log likelihood function

$$\frac{\delta l(\boldsymbol{\beta})}{\delta \boldsymbol{\beta_k}} = \sum_{i=1}^{n} \delta_i \left(x_{ik} - \frac{\sum_{l \epsilon R(y_i)} x_{lk} \exp(\hat{\beta}^t x_l)}{\sum_{l \epsilon R(y_i)} \exp(\beta^t x_l)} \right)$$

Evaluated at $\widehat{\boldsymbol{\beta}}$. The i^{th}term is this expression then corresponds to $r_{s_{ik}}$.

Since

$$\frac{\delta l(\boldsymbol{\beta})}{\delta \boldsymbol{\beta_k}}\Big|_{\widehat{\boldsymbol{\beta}}} = 0$$

The Schoenfeld residuals for each of the parameters sum to zero.

3.24 Checking the proportional hazard assumption

In fitting the Cox model, it is assumed that the hazard ratio between two objects with different covariates information is constant over time. Different techniques have been developed to check whether the proportional hazard assumption holds.

- Anderson (1982) developed a diagnostic plot, in which $\widehat{H}_j(t)$ is plotted against $\widehat{H}_1(t)$. These curves should be a straight line through the origin.

- We plug $log\left(\widehat{H}_1(t)\right), \ldots, log\left(\widehat{H}_g(t)\right)$ against time, the resulting curves should be parallel to one another if the proportional hazards assumption holds.

3.25 Tracking down influential observations

An observation is considered to be influential for a parameter if the perimeter estimate changes considerably when the observation is omitted.

To determine the influence of an observation to be carried out in our study.

We first fit the model with all observations leading to parameter estimates $\hat{\beta}$. Next, we fit the same model leaving observation i out, leading to parameter estimate $\widehat{\boldsymbol{\beta}}_{(-i)}$. The influence of observation i on parameter β_k is then given by $\beta_k - \beta_{k(-i)}$.

An approximation of $\beta_k - \beta_{k(-i)}$ is available through the use of the score residual, which is a function of the Schoenfeld residual.

The score residual is given by

$$r_{sc_{ik}} = r_{s_{ik}} - \sum_{j:y(j)\leq y_i} \left(x_{ik} - \frac{\sum_{l\epsilon R(y_i)} x_{lk} \exp(\hat{\beta}^t x_l)}{\sum_{l\epsilon R(y_i)} \exp(\hat{\beta}^t x_l)} \right) \times \exp(\hat{\beta}^t x_i) \left(\hat{H}_0 (y_{(j)} - y_{(j-1)}) \right) \qquad (3.54)$$

It can be shown that $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_i$ is approximated by

$$\left(I(\hat{\beta}) \right)^{-1} \left(r_{sc_i}, \dots, r_{sc_p} \right)^t$$

Where $I(\hat{\beta})$ is the observed information matrix.

Parametric survival models

In semi-parametric model we did not make any assumption on the distribution of the baseline hazard function $h_0(t)$. Different parametric choices are available and we will discuss them in our study. The most prominent choice is the parametric baseline hazard function that leads to Weibull distributed events times models with this parametric assumption combines different properties. Such models are at the same time proportional hazard models, accelerated failure time models and can be expressed as log-linear models. We will discuss some of presentations.

3.26 Exponential distribution

The baseline hazard function, $h_0(t)$, for the exponentially distributed event times is constant, i.e.,

$$h_0(t) = \lambda$$

From the constant baseline hazard function, it follows that

$$S_0(t) = \exp(-\lambda t)$$

and

$$f_0(t) = \lambda \exp(-\lambda t)$$

the median event time can be obtained by solving the equation $S\left(\frac{1}{2}\right) = 0{,}5$ which leads to $t_{0,5} = log\frac{2}{\lambda}$, and the quantile can be obtained by solving the equation $S(t_p) = 1 - p$, and thus

$$t_p = \frac{-\log(1-p)}{\lambda} \tag{3.55}$$

The main feature of the exponential distribution is thus that the instantaneous hazard does not vary over time. Another property is the lack of memory property.

3.27 Fitting the model

The likelihood is maximized to obtain parameter estimates, both for the case of a single sample and two sample.

3.28 A single sample

The survival likelihood for survival data with events and right censored data is generally given by

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} \left(s(y_i)\right)^{1-\delta_i}$$

which leads for exponentially distributed event times to

$$L = \prod_{i=1}^{n} \left(\lambda exp(-\lambda y_i)\right)^{\delta_i} \left(exp(-\lambda y_i)\right)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \lambda^{\delta_i} exp(-\lambda y_i)$$

And resulting in the log likelihood function

$$l = \sum_{i=1}^{n} \delta_i log\lambda - \lambda \sum_{l=1}^{n} y_i$$

$$= d\ log\lambda - \lambda \sum_{l=1}^{n} y_i$$

with d the total of events differentiating l with respect to λ gives

$$\frac{dl}{d\lambda} = \frac{d}{\lambda} - \sum_{n=1}^{n} i = 1^n y_i$$

and equating this expression to zero leads to the maximum likelihood estimator

$$\hat{\lambda} = \frac{d}{\sum_{i=1}^{n} y_i}$$

The asymptotic variance of $\hat{\lambda}$ can be approximated by the inverse of the observed function. We first derive the second derivative of the log-likelihood function with respect to λ

$$\frac{d^2 l}{d\lambda^2} = \frac{-\lambda^2}{d} \tag{3.56}$$

and the variance is then given by minus the inverse of equation (3.56) evaluated at $\hat{\lambda}$. i.e.,

$$Var(\hat{\lambda}) = d\lambda^2 = \frac{\lambda^2}{d} \tag{3.57}$$

having obtained the variance of $\hat{\lambda}$, we can also derive the variance of the functions of $\hat{\lambda}$, such as the quantiles and the median of the survival function, using the delta method. For example, the approximate variance of the estimated p^{th} quantity is given by

$$Var(\hat{t}_p) \approx \left(\frac{1}{\hat{\lambda}^2}\log(1-p)\right)^2 Var(\hat{\lambda})$$

Now using equations (3.55) and (3.56) obtain

$$Var(\hat{t}_p) \approx \frac{(\hat{t}_p)^2}{d}$$

Again, using the delta method we have

$$Var(log\ \hat{t}_p) \approx \frac{Var(log\ \hat{t}_p)}{(\hat{t}_p)^2}$$

$$\approx \frac{1}{d} \qquad (3.58)$$

after exponentiation the confidence interval for$log\hat{t}_p$, we obtain the $100(1-\alpha)\%$ confidence interval for $\hat{t}_p$ as

$$exp\left(log\hat{t}_p \pm \frac{Z_{\frac{\alpha}{2}}}{d}\right)$$

3.29 Two samples

For two simple we can express the hazard function as

$$h_i(t) = h_0(t)exp(\beta x_i) = \lambda\ exp(\beta x_i)$$

With $x_i = 0$ if subject i belongs to the first group, and $x_i = 1$ if i belongs to the second group.

The likelihood is then given by

$$L = \prod_{i=1}^{n} \left(\lambda exp(\beta x_i)\right)^{\delta_i} \exp\left(-\lambda y_i exp(\beta x_i)\right)$$

Which leads to the log likelihood expression

$$l = d_1 log\lambda + d_2 \log(\lambda\Psi) - \lambda \sum_{i=1}^{n} y_1(1 - x_i) - \lambda\Psi \sum_{i=1}^{n} y_i\, x_i$$

$$= (d_1 + d_2) \log\lambda + d_2 log\, \Psi - \lambda \left(\sum_{i=1}^{n} y_1(1 - x_i) + \Psi \sum_{i=1}^{n} y_i\, x_i \right)$$

With d_1 and d_2 the number of events in the first and second group respectively, $\sum_{i=1}^{n} y_1(1 - x_i)$ and $\sum_{i=1}^{n} y_i\, x_i$ the total at risk time in the first and the second group respectively, and $\Psi = \exp(\beta)$ the hazard ratio of the second versus the first group.

we now take the first derivatives of the log likelihood function with respect to λ and Ψ and equate it to zero

$$\frac{dl}{d\lambda} = \frac{d_1 + d_2}{\lambda} - \sum_{i=1}^{n} y_i(1 - x_i) - \Psi \sum_{i=1}^{n} y_i x_i = 0 \qquad (3.59)$$

$$\frac{dl}{d\lambda} = \frac{d_2}{\Psi} - \lambda \sum_{i=1}^{n} y_i\, x_i = 0 \qquad (3.60)$$

Thus, from equation (3.53), we have

$$\hat{\lambda} = \frac{d_2}{\Psi \sum y_i x_i} \tag{3.61}$$

and substituting this expression (3.56) for $\hat{\lambda}$ in (3.54), we obtain

$$\widehat{\Psi} = \frac{d_2 \sum_{i=1}^{n} y_i (1 - x_i)}{d_1 \sum_{i=1}^{n} y_i x_i} \tag{3.62}$$

now using (3.57) in (3.56) we find

$$\hat{\lambda} = \frac{d_1}{\sum_{i=1}^{n} y_i (1 - x_i)} \tag{3.63}$$

We can obtain an approximation of the asymptotic variance-covariance matrix of $(\hat{\lambda}, \Psi)$ from inverse of the observed information Matrix, which has its entries minus the second derivatives of the log-likelihood function

$$\frac{d^2 l}{d\Psi^2} = \frac{-d_2}{\Psi^2}$$

$$\frac{d^2 l}{d\lambda^2} = \frac{d_1 + d_2}{\lambda}$$

$$\frac{dl}{d\Psi d\lambda} = -\sum yx_i$$

We then have information matrix

$$I = \begin{pmatrix} \frac{d_2}{\Psi^2} & \sum_{i=1}^{n} yx_i \\ \sum_{i=1}^{n} yx_i & \frac{d_1 + d_2}{\lambda^2} \end{pmatrix}$$

with inverse

$$\frac{1}{(d_1 + d_2) - \Psi^2(\sum_{i=1}^{n} yx_i)} \begin{pmatrix} (d_1 + d_2)\Psi^2 & -\Psi^2\lambda^2 \sum_{i=1}^{n} yx_i \\ -\Psi^2\lambda^2 \sum_{i=1}^{n} yx_i & d_2\lambda^2 \end{pmatrix} \tag{3.64}$$

We substitute now the estimator $\hat{\lambda}$ and $\widehat{\Psi}$ in equation (3.59) to obtain the actual approximation for the asymptotic variance-covariance matrix leading to

$$Var(\hat{\lambda}) = \frac{+\hat{\lambda}^2}{d_i}$$

$$Var(\Psi) = \widehat{\Psi}^2 \frac{d_1 + d_2}{d_1 d_2}$$

Since parameters λ and Ψ and cannot be negative, it advisable to derive the confidence intervals based on the logarithm of parameters. The approximate variance of these transformed parameters is given by

$$Var(log\hat{\lambda}) = -\hat{\lambda}^2 Var(\hat{\lambda})$$

$$Var(log\widehat{\Psi}) = -\widehat{\Psi}^2 Var(\Psi)$$

from which follows the $100(1-\alpha)\%$ confidence interval after exponentiation

$$\hat{\lambda} exp\left(\pm Z_{\frac{\alpha}{2}}\sqrt{Var(\log\hat{\lambda})}\right)$$

$$\hat{\Psi} exp\left(\pm Z_{\frac{\alpha}{2}}\sqrt{Var(\log\hat{\lambda})}\right)$$

3.30 The Weibull distribution

the baseline hazard function for Weibull distribution event times is given by

$$h_0 = \lambda p t^{p-1}$$

with λ the scale parameter and p the shape parameter.

it follows that

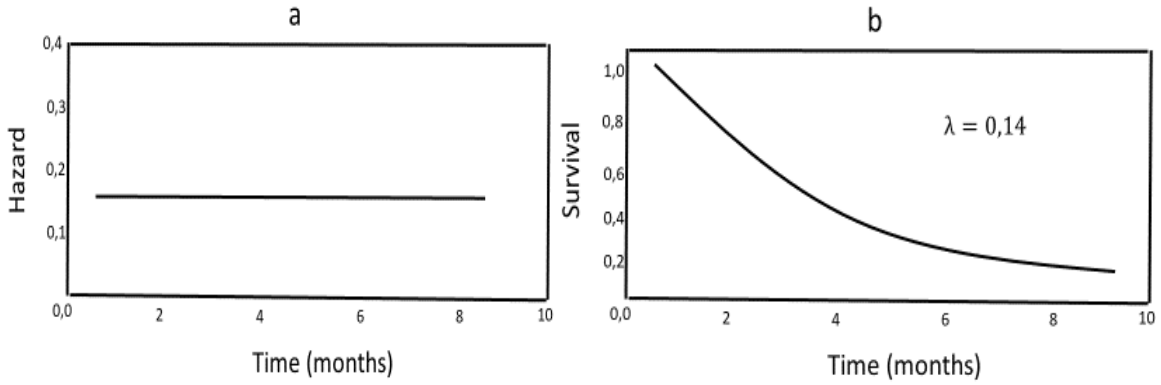
$$S_0(t) = \exp(-\lambda t^p)$$

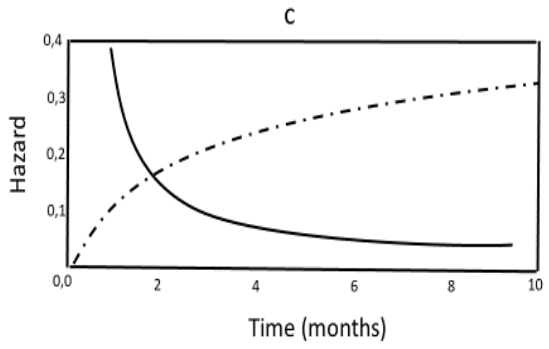
and

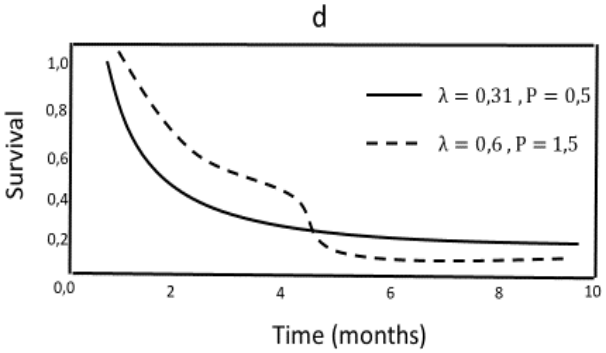
$$f_0(t) = \lambda p t^{p-1} \exp(-\lambda t^p)$$

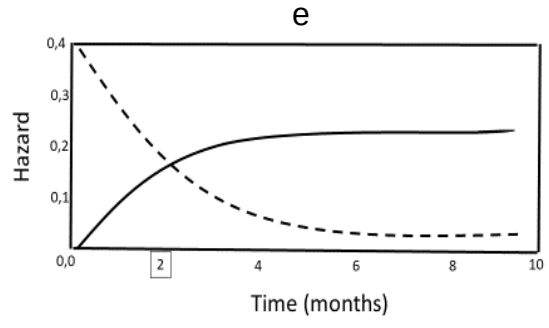
the hazard function for this distribution thus varies with time. If p is less than 1 the hazard decrease monotonically with time, whereas with p greater than 1 the hazard increase monotonically with time. If $p = 1$ the hazard function is constant overtime, and it is equivalent to the exponential distribution. The hazard and survival function for the Weibull distribution with $\lambda = 0{,}4$ and $p = 0{,}6$, both resulting in a median event time of 5, is given in Figure 3.2 Hazard and survival functions for different distributions with, for any parameter setting, a median event time equal to 5. Hazard

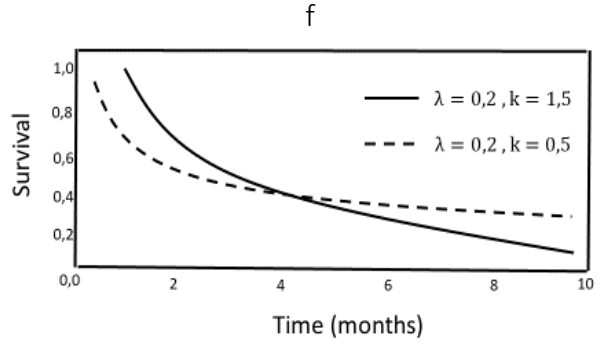
and survival functions for the exponential distribution are given in a and b, for the Weibull in c and d, for the log-logistic distribution in e and f, finally for log-normal distribution in g and h.











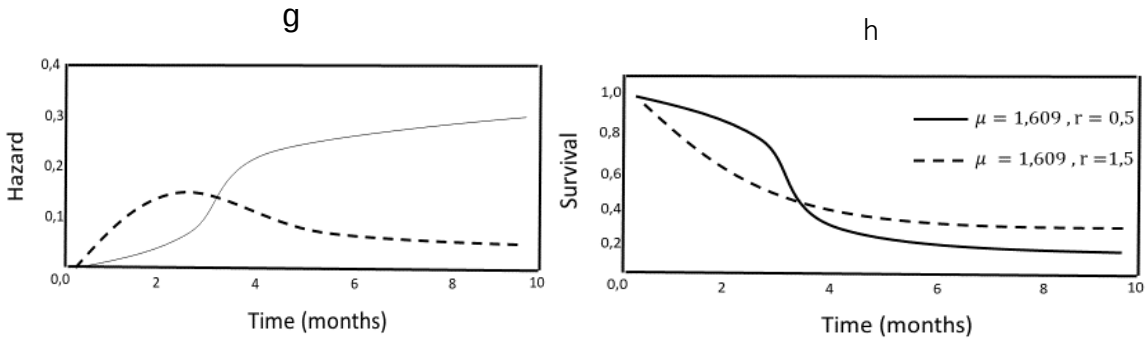


Figure 3. 2: Hazard function h_0, density function $f_0(t)$ and survival function $S_0(t)$ of some distributions used for modelling survival time

In the presence of covariates $\boldsymbol{x}_{i\,,}$ the corresponding hazard function $h_i(t)$ for proportional hazard model is obtained by multiplying the hazard function by $\exp(x_i^t\boldsymbol{\beta})$. From the hazard function, the survival function can be obtained as $exp[-H_i(t)]$ with cumulative hazard $H_i(t) = \int_0^t h_i(s)ds$. The density function is given by $h_i(t)S_i(t)$.

Table 3. 6: Survival Distribution parameters

Distribution	Parameter range	Function	Expression
Exponential	$\lambda > 0$	$h_0(t)$ $f_0(t)$ $S_0(t)$ $Median(T)$	λ $\lambda\exp(-\lambda t)$ $\exp(-\lambda t)$ $\frac{log2}{\lambda}$
Weibull	$\rho, \lambda > 0$	$h_0(t)$ $f_0(t)$ $S_0(t)$ $Median(T)$	$\lambda\rho t^{\rho-1}$ $\rho\lambda t^{\rho-1}\exp(-\lambda t^{\rho})$ $\exp(-\lambda t^{\rho})$ $\left(\frac{log2}{\lambda}\right)^{\frac{1}{\rho}}$
Log-logistic	$\lambda, k > 0$	$h_0(t)$ $f_0(t)$ $S_0(t)$ $Median(T)$	$\frac{kt^{k-1}\lambda^k}{1+(t\lambda)^k}$ $\frac{k^{t-1}\lambda^k}{[1+(t\lambda)^k]^2}$ $\frac{1}{1+(t\lambda)^k}$ $exp\left(\frac{1}{\lambda}\right)$
Log-normal	$\mu\epsilon R, r > 0$	$h_0(t)$ $f_0(t)$ $S_0(t)$ $Median(T)$	$\frac{f(t)}{S(t)}$ $\frac{1}{t\sqrt{2\pi r}} exp\left[1\frac{1}{2r}(log(t)-\mu)^2\right]$ $1 - F_N\left(\frac{log(t)-\mu}{\sqrt{r}}\right)$ $\exp(\mu)$
Gamma	$\alpha, \lambda > 0$	$h_0(t)$ $f_0(t)$ $S_0(t)$	$\frac{\lambda(\lambda t)^{\alpha-1}e^{-\lambda t}}{\Gamma(\alpha)\,\gamma(\alpha,\lambda t)}$ $\frac{\lambda(\lambda t)^{\alpha-1}e^{-\lambda t}}{\Gamma(\alpha)}$ $1 - \frac{\gamma(\alpha,\lambda t)}{\Gamma(\alpha)}$

The median event time can be obtained by solving the equation $S(t_{0,5})$=0,5 which leads to

$$t_{0,5} = \left(\frac{log2}{\lambda}\right)^{\frac{1}{\rho}}$$

more generally, the p^{th} can be obtained by solving the equation

$$S(t_p) = 1 - \rho$$

and this

$$t_p = \left(\frac{-\log(1-\rho)}{\lambda}\right)^{\frac{1}{\rho}} \tag{3.65}$$

an imperial check for the Weibull distribution is provided by plotting the log cumulative hazard (i.e., $logH(t)$ as a function of log time). This will yield a straight line as

$logH(t) = log\ \lambda + \rho\ log\ t$ if the Weibull assumption holds.

3.30.1 Fitting the model

The survival likelihood for Weibull distributed survival data with event times and write censored data is generally given by

$$L = \prod_{i=1}^{n} \left(\lambda\rho y_i^{\rho-1}\exp\left(-\lambda y_i^{\rho}\right)^{\delta_i}\left(exp\left(-\lambda y_i^{\rho}\right)\right)^{1-\delta_i}\right)$$

resulting in the log likelihood function

$$l = \sum_{i=1}^{n} d\ log(\lambda\rho) + (\rho - 1)\sum_{i=1}^{n} \delta_i\ log\ y_i - \lambda \sum_{i=1}^{n} y_i^{\rho}$$

with d the total number of events. Maximum likelihood estimator can be obtained by equalling the first derivatives of l with respect to to λ and p to zero

we have

$$\frac{dl}{d\lambda} = \frac{d}{\lambda} - \sum i = 1^{n} y_i^{\rho} = 0 \tag{3.66}$$

$$\frac{dl}{d\lambda} = \frac{d}{\rho} + \sum_{i=1}^{n} \delta_i\ logy_i - \lambda \sum_{i=1}^{n} y_i^{\rho}\ logy_i = 0 \tag{3.67}$$

and thus we have from (3.66) that

$$\hat{\lambda} = \frac{d}{\sum_i = 1^{n} y_i^{\rho}} \tag{3.68}$$

and substituting (3.70) in (3.67) we obtain

$$\frac{d}{\hat{\rho}} + \sum_{i=1}^{n} \delta_i logy_i - \frac{d}{\sum_{i=1}^{n} y_i^{\rho}} \sum_{i=1}^{n} y_i logy_i = 0 \tag{3.69}$$

This equation (3.69) is nonlinear in $\hat{p}$ and can only be solved by numerical procedure such as Newton-Raphson algorithm. We can also obtain an approximation of the asymptotic variance-covariance matrix by taking, the inverse of the observed information matrix.

we can make use of the following extension of the Taylor series expansion approximation for a function of two variables $\theta_1$and θ_2, $g(\theta_1, \theta_2)$.

The variance of g can be approximated by

$$Var(g) = \left(\frac{\delta g}{delta\theta_1}\right)^2 Var(\hat{\theta}_1) + \left(\frac{\delta y}{delta\theta_2}\right)^2 Var(\hat{\theta}_2) + 2\left(\frac{\delta g}{delta\theta_1}\frac{\delta g}{delta\theta_2}\right) Cov(\theta_1, \theta_2) \quad (3.70)$$

we first obtain the variance of $log\hat{t}_p$

$$log\hat{t}_p = \frac{1}{\hat{\rho}} log\left(\frac{-\log(1-\rho)}{\hat{\lambda}}\right)$$

Now applying the approximation (3.63) to $log\hat{t}_p$ we obtain

$$Var\left(log\hat{t}_p\right) = \left(\frac{\delta log\hat{t}_p}{delta\ \lambda}\right)^2 Var(\hat{\lambda}) + \left(\frac{\delta log\hat{t}_p}{delta\ p}\right)^2 Var(\hat{\rho}) + 2\left(\frac{\delta log\hat{t}_p\ \delta log\hat{t}_p}{delta\ \lambda\ delta\ \rho}\right) Cov(\hat{\lambda}, \hat{\rho}) \quad (3.71)$$

The derivatives are given by

$$\frac{dlog\hat{t}_p}{d\lambda} = \frac{-1}{\lambda\rho}$$

$$\frac{dlog\hat{t}_p}{d\rho} = \frac{-\log\left(log(1-\rho) - log(\lambda)\right)}{\rho^2}$$

Having obtained the variance of $\hat{\lambda}$, we can also derive the variance of the function of $\hat{\lambda}$ such as the quantiles and the median of the survival function, using the delta method. For instance, the approximate variance of the estimate p^{th} quantile is given by

$$Var\left(log\hat{t}_p\right) \approx \frac{Var(\lambda)}{\lambda^2\hat{\rho}^2}\lambda^2 \frac{\left(log\left(log(1-\rho)\right) - log\left(\hat{\lambda}\right)\right)^2}{\hat{\rho}^4} Var(\hat{\rho}) + 2\frac{log\left(log(1-p)\right)}{\hat{\lambda}\,\hat{p}^3} Cov\left(\hat{\lambda},\hat{\rho}\right)$$

We will obtain the $100(1-\alpha)\%$ confidence Interval for $log\hat{t}_p$ and then exponentiate to obtain the confidence interval for $\hat{t}_p$

$$\hat{t}_p exp\left(\pm Z_{\frac{\alpha}{2}}\sqrt{Var(log\hat{t}_p)}\right)$$

3.30.2 The general Weibull baseline hazard model

The Weibull model can be generally presented as

$$h_i(t) = h_0(t)\exp(\beta^t x_i) \tag{3.72}$$

With $h_0(t) = \lambda p t^{\rho-1}$and $\boldsymbol{\beta}$ a ρ ×1 vector containing the parameters.

The event time of the i^{th} subject is then characterized by the Weibull distribution with scale parameter $\lambda\exp(\boldsymbol{\beta}^t\boldsymbol{x}_i)$ and shape parameter p. Thus, all subjects share the shape perimeter but differ with respect to their scale parameter. We have the corresponding survival function

$$S_i(t) = exp(-\lambda\exp(\boldsymbol{\beta}^t\boldsymbol{x}_i)t^{\rho})$$

and density function

$$f_i(t) = \lambda\rho t^{\rho-1}\exp(\boldsymbol{\beta}^t\boldsymbol{x}_i)\exp\left(-\lambda exp(\boldsymbol{\beta}^t\boldsymbol{x}_i)\right)t^{\rho}$$

using these density and survival functions, the likelihood function can be constructed and then maximized for λ, ρ and $\boldsymbol{\beta}$. Such maximization is typically based on the Newton-Raphson algorithm, which also provides an approximation for the asymptotic variance-covariance matrix of the parameters.

3.31 The log-linear model representation

The general model (3.72) can also be represented alternatively as a log-linear model. For the randomly distributed event times T_i, the log-linear model is given by

$$logT_i = \mu + \boldsymbol{\alpha}^t \boldsymbol{x}_i + \delta E_i \tag{3.73}$$

Where $\boldsymbol{\alpha}$ is a vector containing the parameters, μ is the intercept and δ is the scale parameter of the log-linear model. E_i is a random error term which describes the division of $logT_i$ from the prescribed value $\mu + \boldsymbol{\alpha}^t \boldsymbol{x}_i$. If we assume that E_i has a Gumbel distribution

$$f_E(e) = \exp\big(e - exp(e)\big)$$

With

$-\infty < e < \infty$, then the log-linear model (3.66) results in Weibull distributed event times as we will now demonstrate as E_i has a Gumbel distribution, the transformed variable

$Z_i = \exp(E_i)$ has a

$$f_Z(z) = \frac{\exp\big(logz - exp(logz)\big)}{z}$$

$$= exp(-z)$$

which corresponds to an exponential distribution with mean one.

We now consider the survival function using the log-linear model (3.73)

$$\begin{aligned} S_i(t) &= P(T_i \geq t) \\ & P(logT_i \geq logt) \\ &= P(\mu + \boldsymbol{\alpha}^t \boldsymbol{x}_i + \delta E_i \geq t) \\ &= P\left((E_i) \geq \left(\frac{logt - \mu - \boldsymbol{\alpha}^t \boldsymbol{x}_i}{\delta}\right)\right) \end{aligned} \tag{3.74}$$

We have shown above that $\exp(E_i)$ has a unit mean exponential distribution so that

$$P(\exp(E_i) \geq Z) = \exp(-Z)$$

applying this to (3.74) we have

$$S_i(t) = exp\left(-exp\left(\frac{logt - \mu - \boldsymbol{\alpha}^t \boldsymbol{x}_i}{\delta}\right)\right) \tag{3.75}$$

We now compare the survival expression (3.74) to the survival expression based on the hazard model (3.72), i.e.,

$$S_i(t) = exp(-exp(\boldsymbol{\beta}^t \boldsymbol{x}_i)\lambda t^p) \tag{3.76}$$

Clearly, we can see that the survival functions (3.75) and (3.76) are the same but

$$\lambda = exp\left(\frac{-\mu}{\delta}\right)$$

$$\rho = \delta^{-1}$$

$$\beta_j = \frac{-\alpha_j}{\sigma}$$

Therefore, having estimated for μ,δ and σ, estimate for λ, ρ and β can be immediately obtained. Obtaining the variance of λ, ρ and β from the variances of μ, σ and $\boldsymbol{\alpha}$ is not straight forward. We will demonstrate how that can be done for one of the components of the parameter vector of interest $\boldsymbol{\beta}$. We need to derive the variance of a ratio of two parameter estimates

$$Var(\hat{\beta}_j) = \frac{-\hat{\alpha}_j}{\hat{\sigma}}$$

This can be approximated by using the Delta method.

The Delta method start from the one-term Taylor expansion of

$g(\hat{\varsigma})$, with $\hat{\varsigma}$ the maximum likelihood estimator of $\hat{\varsigma}$, (note ς is the Greek letter for sigma variant)

$$g(\hat{\varsigma}) \approx g(\varsigma) + r^t(\hat{\varsigma} - \varsigma) \tag{3.77}$$

With

$$r^t = \left(\frac{\partial g(\varsigma)}{\partial(\varsigma_1)}, \ldots\ldots, \frac{\partial g(\varsigma)}{\partial \varsigma_k}\right)$$

The vector of the first partial derivatives evaluated at ς (sigma variant).

from (3.77) we obtain

$$Var\big(g(\hat{\varsigma})\big) \approx r^t Var(\hat{\varsigma} - \varsigma) \approx r^t Var(\hat{\varsigma}) r$$

Where in the last approximation we ignore the bias contribution. $Var(\hat{\varsigma})$ Is the variance-covariance of matrix of $\hat{\varsigma}$ (sigma variant).

For the specific case of Weibull proportional hazard model, we have

$\varsigma^t = (\mu, \boldsymbol{\alpha}, \sigma)$ and

$$\hat{\beta}_j = \frac{-\hat{\alpha}_j}{\hat{\sigma}}$$

We therefore, have

$$r^t = \left(0,0, \dots, \frac{-1}{r}, \dots 0, \frac{\alpha_j}{\sigma^2}\right)$$

Given many zeros it is easily observed that

$$Var\big(\hat{B}_j\big) \approx r^t Var(\hat{\varsigma}) r$$

$$= \begin{pmatrix} -\frac{1}{\sigma} & \frac{\alpha_j}{\sigma} \end{pmatrix} \begin{pmatrix} Var(\hat{\alpha}_j) & Cov(\hat{\alpha}_j, \hat{\sigma}) \\ Cov(\hat{\alpha}, \hat{\sigma}) & Var(\hat{\sigma}) \end{pmatrix} \begin{pmatrix} -\frac{1}{\sigma} & \frac{\alpha_j}{\sigma^2} \end{pmatrix}^t$$

$$= \frac{1}{\sigma^2} Var\big(\hat{\alpha}_j\big) - 2\frac{\alpha_j}{\sigma^3} Cov\big(\hat{\alpha}_j, \hat{\sigma}\big) + \frac{{\alpha_j}^2}{\sigma^4} Var(\hat{\sigma}) \qquad (3.78)$$

an estimated for $Var\big(\hat{B}_j\big)$ is obtained by using in (3.78) as estimates for $Var\big(\hat{\alpha}_j\big), Var\big(\hat{\sigma}_j\big)$ and $Cov\big(\hat{\alpha}_j, \hat{\sigma}\big)$ the corresponding entries of the inverse of the

observed information matrix and by replacing in (3.78) α_j and σ by their corresponding estimates $\hat{\alpha}_j$ and $\hat{\sigma}$ obtained by fitting the log-linear model.

The accelerated failure time (AFT) model

Another representation of Weibull distributed event times consist of the accelerated time model. The Accelerated failure time model is an alternative if the proportional hazard assumption does not hold. In contrast to the proportional hazard model, the accelerated failure time model is the best characterized in terms of the survival function with $S_T(t)$ and $S_c(t)$ the survival functions in the treated and control population, the AFT model specifies that, with $\Phi > 0$, $S_T(t) = S_c(\Phi t)$

The interpretation is as follows:
The percentage of subjects in the treatment group that lives longer than time t equals the percentage of subjects in the control group that lives longer than $\Phi(t)$. The parameter Φ is called the acceleration factor, values below one are in favour of the treatment, as the survival time is that prolonged under the treatment, an alternative interpretation is in terms of the median survival times. With M_T and M_c the median survival times in the treatment and control group, we have

$$S_T(M_T) = S_C(M_C) = 0{,}5$$

From the accelerated failure time assumption, it follows that

$$S_T(M_T) = S_c(\Phi M_T) = 0{,}5$$

and therefore $\Phi M_T = M_C$. For $\Phi < 1$
the median survival time in the treatment group is larger than the median survival in the controlled group. This demonstrated in Figure 3.3

Figure 3.3: Survival function of the control group with Weibull distributed event times ($\lambda = 2$ and $\rho = 2$) and of the treatment group with acceleration factor in the accelerated failure time model. The accelerated time model is equal to 0,5. M_C and M_T are the median event times for the control and treated groups

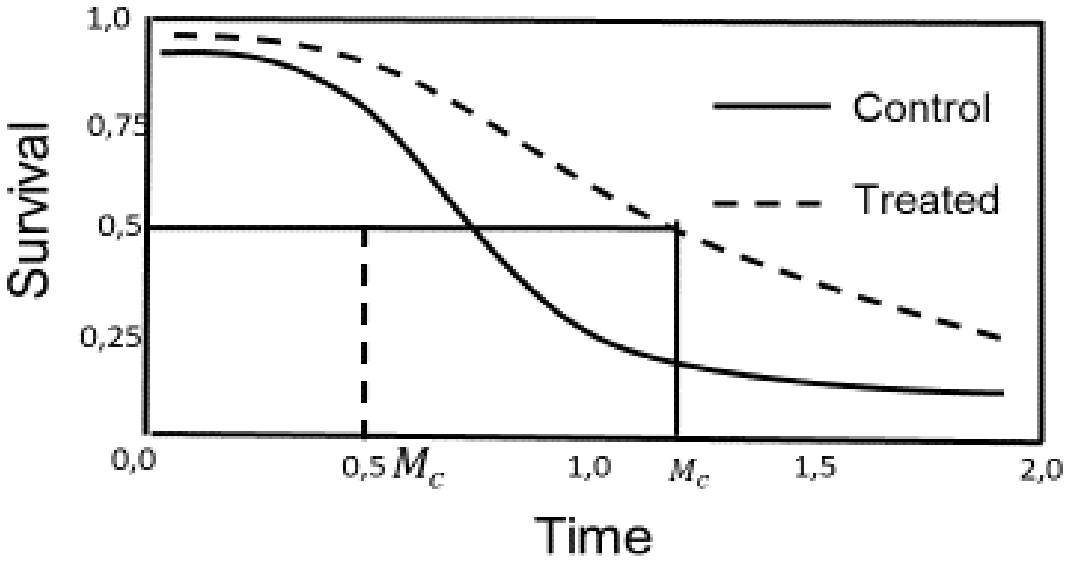


Figure 3. 3: Survival function of the control group with Weibull distributed event times

From the expression of the accelerated failure time model in terms of survival, it follows that

$f_T(t) = \Phi f_c(\Phi t)$ and

$$h_T(t) = \Phi h_c(\Phi t)$$

since $\Phi > 0$, it is convenient to set $\Phi = \exp(\beta)$, $\beta \epsilon R$, leading to

$$h_T(t) = \exp(\beta)\, h_c(exp(\boldsymbol{x_i^t \beta})t)$$

For general covariates, the accelerated failure time model can be written as

$$h_i(t) = exp(\boldsymbol{x_i^t\beta})h_0(exp(\boldsymbol{x_i^t\beta})t) \quad (3.79)$$

Again we will assume a particular parametric form for the baseline hazard function $h_0(t)$. At first proposal is to use the same parametric form as in the parametric proportional hazards model, $h_0(t) = \lambda\rho t^{\rho-1}$. However, the argument of the baseline hazard in (3.79) differs from the proportional hazards model as it also contains the covariate information $exp(\boldsymbol{x_i^t\beta})$. When making this parametric assumption for the baseline hazard, it again follows that the event times are Weibull distribution.

The hazard for this parametric model can be written as

$$h_i(t) = \left(exp(\boldsymbol{x_i^t\beta})\right)^{\rho}\lambda\,\rho\,t^{\rho-1} \quad (3.79)$$

It follows that

$$S_i(t) = exp\left(-\lambda t^{\rho} exp(\boldsymbol{\rho x_i^t\beta})\right) \quad (3.80)$$

and

$$f_i(t) = \lambda\,\rho\,t^{\rho-1}\,exp(\rho\boldsymbol{x_i^t\beta})exp\left(-\lambda t^{\rho-1}exp(\rho\boldsymbol{x_i^t\beta})\right)$$

and therefore

$$T_i \sim W\left(\lambda exp(\rho\boldsymbol{x_i^t\beta}, \rho)\right)$$

Thus, all subjects have Weibull distribution event times with the same shape parameter ρ but differ with respect to the scale parameter.

3.32 The Gamma distribution

The Gamma distribution is another extension of the exponential distribution commonly used in survival analysis. The Gamma distribution is characterized by two parameters: shape (α) and rate (β), both of which must be greater than zero. The Gamma distribution is particularly flexible, as it allows for a variety of hazard shapes depending on the values of its parameters. When α = 1, the Gamma distribution reduces to the exponential distribution, representing constant hazard over time. As α increases, the hazard shape becomes more peaked, indicating higher initial hazard rates that decrease over time.

The Gamma distribution can model both increasing and decreasing hazard rates. When α < 1, the hazard decreases over time, making it useful for situations where the risk of an event diminishes over time. Conversely, for α > 1, the hazard increases initially before declining, capturing scenarios where the risk initially rises before stabilizing or decreasing.

3.33 The log-logistic distribution

In addition to the above-mentioned distributions, another flexible and popular distribution for time-to-event data is the log-logistic. A random variable X has a log-logistic distribution if its logarithm follows the logistic distribution. It is characterised by two parameters λ and K. The hazard function for the log-logistic distribution with $\lambda = 0{,}2$ and $\lambda = 1{,}5$ and λ=0,2 and $\lambda = 1{,}5$ both resulting in a median event time equal to 5, is given in Figure 3.1 g and 3.1 h

3.34 The Log normal distribution

The log-normal distribution the Log-logistic but is mathematically tractable. A random variable X follows a log normal-distribution if its logarithm follows a normal distribution. $\left(\text{i. e.}, logx \sim N(\mu, \sigma^2)\right)$. It is characterised by two parameters μ and r. The hazard and survival function for the log-normal distribution with $\mu = 1{,}609438$ and $r = 0{,}5$ and $\mu = 1{,}609438$ and $r = 1{,}5$, both resulting in a median event equal to 5, is given in figure 3.1 i and 3.1 j

3.35 Comparison of Cox PH and Parametric models

The proportional models are routinely employed for analysis of time-to-event data in medical research in the presence of covariates (Cox & Snell, 1968), however, parametric models may offer advantages. If the assumption of proportional hazard is violated, then, the results from a PH model will be difficult to generalise to situations where the length of follow-up which differs to that used in the analysis. Furthermore, it would be difficult to translate the results into the effect upon the expected median duration of illness for a patient in a clinical setting (Patel *et al.*, 2006). The second disadvantage of the PH model is that the underlying hazard function is common across all patients, for example, the hazard function for any two patients with baseline h_0 vector $\boldsymbol{x_1}$ and $\boldsymbol{x}_2$ are constrained to be proportional and the method of estimation is based on that , whereas, the method of estimation depends critically on evolving risk sets through time (Kay, 2002). The PH model is considered to be semi-parametric and as such has advantage of being able to cope with variety of basic shapes for the common hazard functions across patients.

The accelerated failure time approach is an alternative strategy for the analysis of time-to-event data and can be suitable even when hazards are not proportional and

this family of models contains a certain form of PH as a special case. The results of AFT model may be easier to interpret and more relevant to clinicians, as they can be directly translated into expected reduction or prolongation of median time to event, unlike the hazard ratio. Based on asymptotic results, Efron (1977) and Oakes (1977) showed that, under certain circumstances, AFT approach leads to more efficient parameter estimates than PH models. With decreasing sample size, relative efficiencies may further change in favour of AFT models. When empirical information is sufficient, AFT model can provide some insight into the shape of the baseline hazard. Furthermore, extrapolation of survival functions becomes possible (Gelber *et al.,*1993)

Chapter 4

Results and discussions

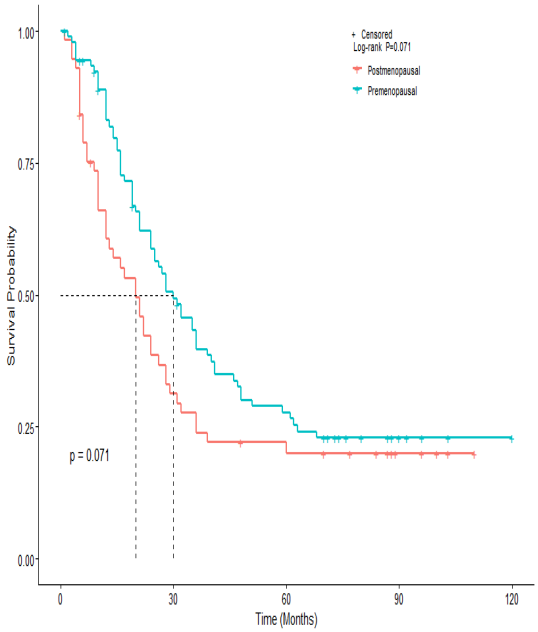
4.1 Introduction

This chapter presents the analysis and discussion of breast cancer women data obtained from Pietersburg Oncology Hospital, in Limpopo Province, South Africa, This chapter consists of three analytic parts, namely Kaplan-Meier curves, log-logistic model and Cox Proportional model. The first section presents Kaplan-Meier curves with different factors that contribute to poor or better prognosis. The second section focuses on the analysis and discussion of factors (menopause, oestrogen etc.) that contribute to poor or better prognosis of breast cancer women using log-logistic model. The final section analyses the factors using Cox Proportional hazard model.

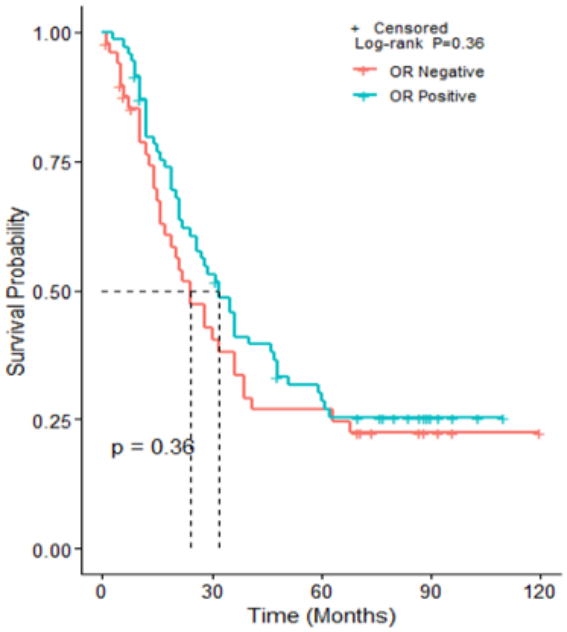
4.2 Analysis and discussion with Kaplan-Meier curves

The survival curves for BC patients in Pietersburg Hospital are presented in Figure 4.3. The plots displayed distinct separation between Kaplan-Meier curves, reflecting strong negative effects on the main effects, that is, menopausal status, oestrogen receptor, progesterone receptor, HER2, CK5/6, Grades, T-Stages, N-Stages, M-Stages, AJCC stages and age at baseline.

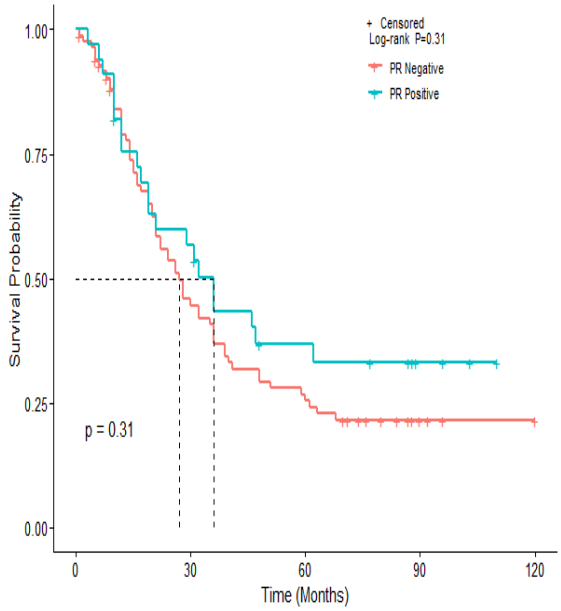
In estimating these survival curves in univariate Cox Proportional hazard model, one covariate at the time was fitted in the model while other covariates were fixed at zero or held constant; for example, when a K-M curve for menopausal status covariate was plotted, covariates such as oestrogen receptor, progesterone receptor, HER 2, CK/6, Grades, T-Stages, N-Stages, M-Stages, AJCC, age at baseline were all fixed at zero, so that survival curves for premenopausal and postmenopausal can be compared effectively (Fox, 2016).



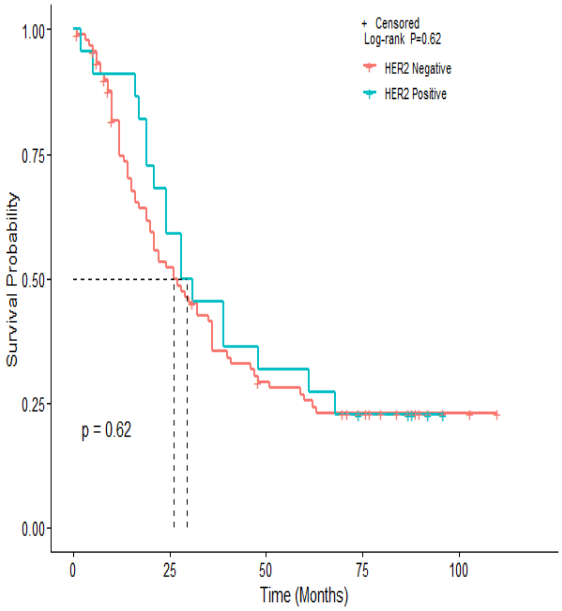
a) Kaplan-Meier survival curves by menopause status



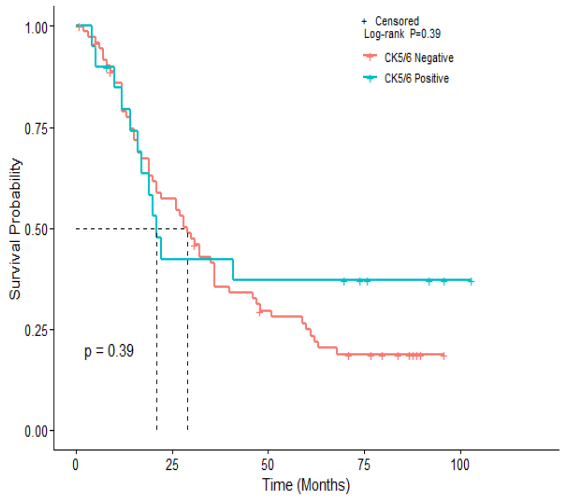
b) Kaplan-Meier survival curves by oestrogen receptor



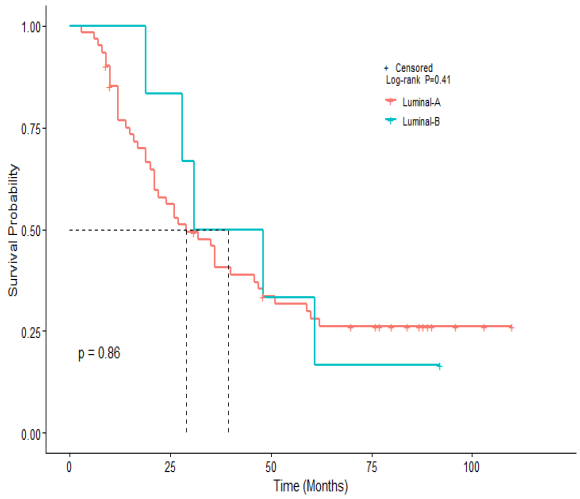
c) Kaplan-Meier survival curves by progesterone receptors



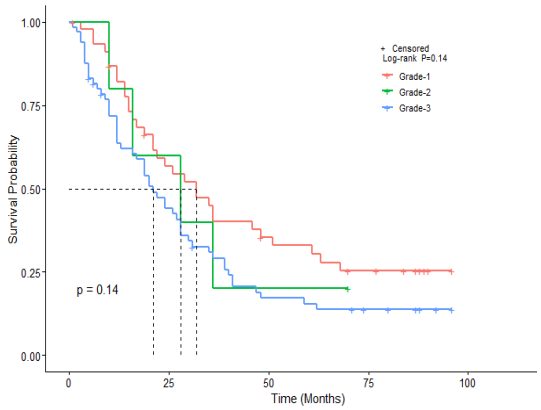
d) Kaplan-Meier survival curves by HER2



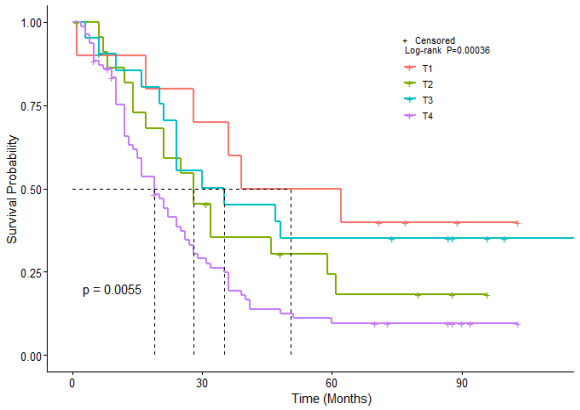
e) Kaplan-Meier survival curves by CK 5\6



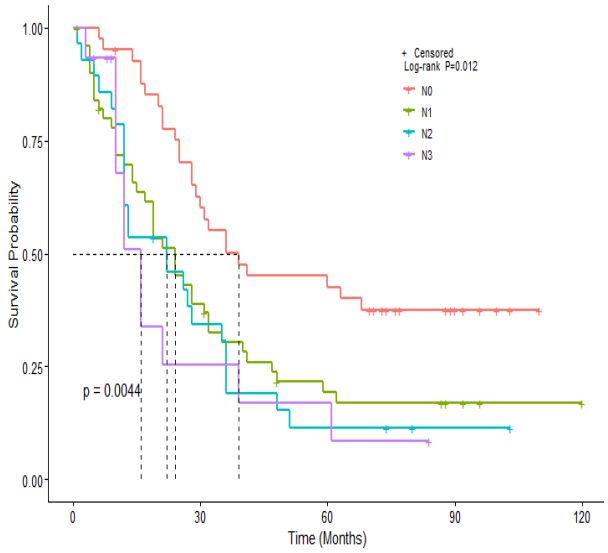
f) Kaplan-Meier survival curves by Luminal subtype



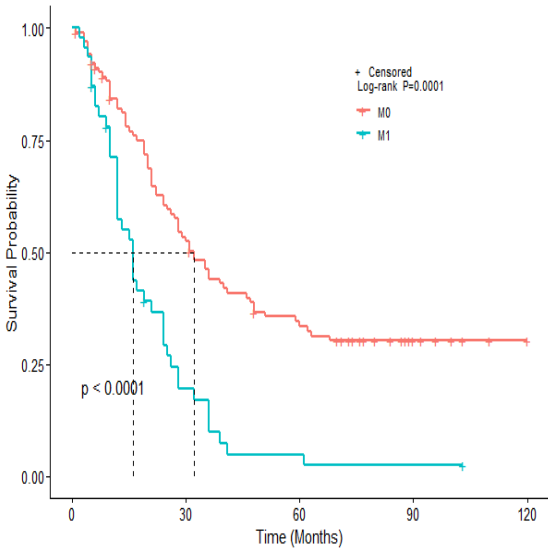
g) Kaplan-Meier survival curves by Tumour grade.



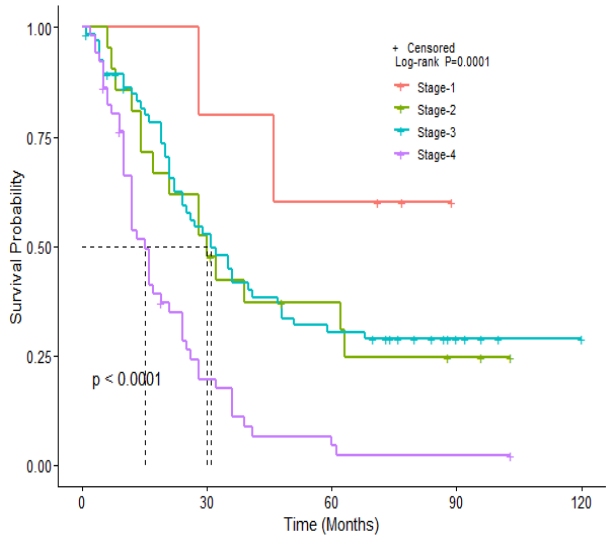
h) Kaplan-Meier survival curves by T-stage



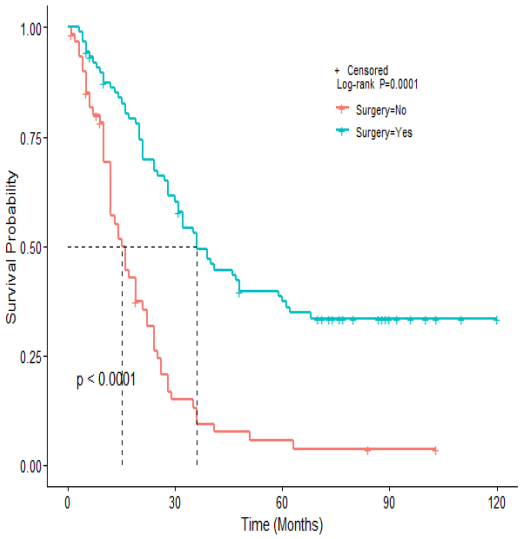
i) Kaplan-Meier survival curves by N-stage



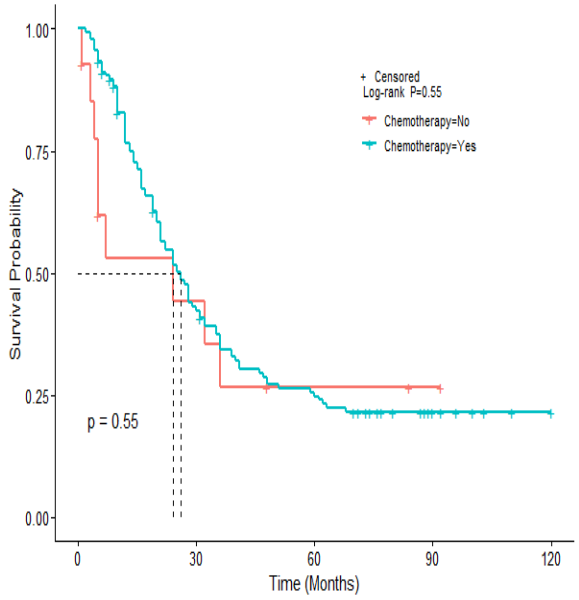
j) Kaplan-Meier survival curves by M- stage



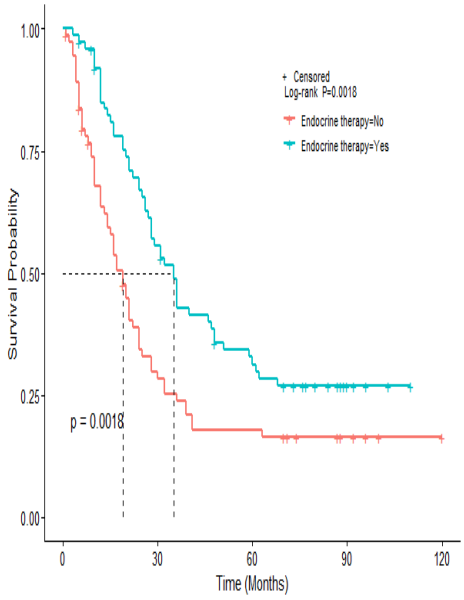
k) Kaplan-Meier survival curves by AJCC group stages



l) Kaplan-Meier survival curves by Surgery



m) Kaplan-Meier survival curves by Chemotherapy



n) Kaplan-Meier survival curves by Endocrine therapy

Figure 4. 1: Kaplan-Meier curves by different covariates

Table 4. 1 Summary of Menopausal status

	n	events	median	0.95LCL	0.95UCL
Postmenopause	57	44	20	12	28
Premenopause	92	66	30	24	40

Table 4. 2 Summary of Oestrogen receptor

	n	events	median	0.95LCL	0.95UCL
OR=Negative	49	35	24	17	39
OR=Positive	70	50	32	24	48

Table 4. 3 Summary of progesterone receptor

	n	events	median	0.95LCL	0.95UCL
PR=Negative	84	62	27	21	36
PR=Positive	33	21	36	19	NA

Table 4. 4 Summary of human epidermal growth factor receptor 2(HER2)

	n	events	median	0.95LCL	0.95UCL
HER2_=Negative	91	65	26.0	20	36
HER2_=Positive	22	17	29.5	24	68

Table 4. 5 Summary of Cytokeratin 5/6

	n	events	median	0.95LCL	0.95UCL
CK56_=Negative	73	56	21	21	36
CK56_=Positive	20	12	29	17	NA

Table 4. 6 Summary of Luminal Subtypes

	n	events	median	0.95LCL	0.95UCL
Subtype_=Luminal-A	61	43	29.0	21	47
Subtype_=Luminal-B	6	5	39.5	28	NA

Table 4. 7 Summary of Tumour Grade

	n	events	median	0.95LCL	0.95UCL
Grade_=Grade 1	46	32	32	21	61
Grade_=Grade-2	5	4	28	16	NA
Grade_=Grade-3	65	53	21	16	30

Table 4. 8 Summary of T stages

	n	events	median	0.95LCL	0.95UCL
T_stage_=T1	10	6	50.5	28	NA
T_stage_=T2	23	17	28.0	21	61
T_stage_=T3	21	13	35.0	24	NA
T_stage_=T4	78	67	19.0	15	26

Table 4. 9 Summary of N stages

	n	events	median	0.95LCL	0.95UCL
N_stage_=N0	41	25	39	29	NA
N_stage_=N1	51	40	24	17	32
N_stage_=N2	28	24	22	12	36
N_stage_=N3	15	11	16	10	NA

Table 4. 10 Summary of M-stages

	n	events	median	0.95LCL	0.95UCL
M_stage_=M0	103	68	32	26	47
M_stage_=M1	46	42	16	12	24

Table 4. 11 Summary of AJCC stages

	n	events	median	0.95LCL	0.95UCL
AJCC_stage_=Stage-1	5	2	NA	46	NA
AJCC_stage_=Stage-2	21	15	30	17	NA
AJCC_stage_=Stage-3	67	45	31	24	48
AJCC_stage_=Stage-4	51	47	15	12	24

Table 4. 12 Summary of Patient Surgery

	n	events	median	0.95LCL	0.95UCL
Surgery_=No	61	54	15	12	22
Surgery_=Yes	88	56	36	31	60

Table 4. 13 Summary of Chemotherapy treatment

	n	events	median	0.95LCL	0.95UCL
Chemotherapy_=No	14	9	24	5	NA
Chemotherapy_=Yes	135	101	26	21	32

Table 4. 14 Summary of Endocrine therapy treatment

	n	events	median	0.95LCL	0.95UCL
ET_=No	74	58	19	14	24
ET_=Yes	75	52	35	28	48

From Sub-figure 4.3a and Table 4.3 illustrate that the Kaplan-Meier curve for postmenopausal breast cancer patients indicates a lower survival probability compared to premenopausal patients. The findings from K-M curves of our study can inform Physicians to take the correct treatment decisions such as choices of adjuvant therapies for pre-menopausal and post-menopausal patients. Table 4.3 shows that the median survival time, the time at which 50% of patients have experienced death, is more than 30 months for pre-menopausal patients, whereas it is 20 months for post-menopausal patients, thus indicates that pre-menopausal patients tent to survive longer.

Sub-Figure 4.3 b and Table 4.4. The Kaplan-Meier curve for positive oestrogen receptor (ER-positive) patients remains higher before 60 months after diagnosis, indicating a higher probability of survival compared to negative oestrogen receptor (ER-negative). Crossing of the curves occurs approximately 60 months after diagnosis suggests that the crossing is not statistically meaningful or clinically relevant. The median survival time for ER-positive is 32 months compared to ER-negative which is 24 months after diagnosis. The ER-positive patients have a better prognosis compared to ER-negative patients. This is because ER-positive cancers can be treated with hormonal therapy, which can help slow the growth of the cancer.

Sub-Figure 4.3 c and Table 4.5 In the beginning, and before 30 months after diagnosis, K-M curves of progesterone receptors kept on crossing each other which makes the interpretation difficult. However, after 30 months, the positive progesterone receptor (PR-positive) has a higher survival probability compared to negative progesterone receptor (PR-negative). This is because PR-positive cancers can also be treated with hormonal therapy, which can help slow the growth of the cancer. According to Table 4.5, the median survival of PR-positive is 36 months and that of PR-negative is 27 months after 30 months. The potential impact of PR status on survival is crucial in helping patients and their healthcare providers make informed decisions about treatment and care.

Sub-Figure 4.3 d and Table 4.6. The K-M curves of human epidermal growth receptor factor 2 (HER2) after diagnosis cross each other at one point in the beginning and thereafter, the HER2 positive has a higher survival probability than HER2 negative up and including 70 months. The HER2-positive breast cancer patients often had a poorer prognosis compared to HER2-negative patients. However, with the development of drugs like Herceptin (trastuzumab) and being applied the survival rates for HER2-positive patients have significantly improved. Table 4.6 shows that the median survival probability of HER2 positive is approximately 30 months compared with 26 months of HER2 negative. The HER2-negative patients generally have a better prognosis, but other factors such as the stage of the cancer and the type of treatment received can also influence survival.

Sub-Figure 4.3 e and Table 4.7: The K-M curves of cytokeratin 5/6 BC patients before 30 months after diagnosis were crossing each other, which suggests that they are clinically irrelevant, however, after 30 months K-M curves show that CK 5/6 positive patients have a higher survival probability than CK 5/6 negative. The CK5/6-positive breast cancers may have a better prognosis compared to those with CK5/6 negative cancers. The CK5/6 status will help Physicians to take an informed treatment decision and different therapeutical approaches. From Table 4.7, the

median survival probability of CK 5/6 negative patients is 21 months and CK 5/6 is 29 months for positive.

Sub-Figure 4.3 f and Table 4.8: The K-M curves show that luminal subtype B breast cancer curve tends to stay higher for a longer period. Clearly, patients with luminal B breast cancer have a better chance of survival compared to luminal A. The luminal subtype A is a significant prognostic factor, however, other factors such as HER2 and triple-negative and treatment can influence survival. These factors may be considered in conjunction with the luminal subtype to provide a far more comprehensive understanding of prognosis. In Table 4.8, the median of luminal-B is approximately 40 months and of luminal-A is 29 months, and this suggest that the majority of BC patients in this luminal subtype B exhibit better survival, whereas luminal subtype A is worse.

Sub-Figure 4.3 g and Table 4.9: Tumour grade-3 breast cancer patients have the lowest K-M curve, suggesting the worst prognosis, while tumour grade-1 is above grade 2 and grade 3 K-M curves, indicating the better prognosis. Tumour grade-2 falls somewhere in between. The K-M curve of tumour grade-2 keeps on crossing tumour grade-1 and grade-3, it implies that tumour grade 2 diagnosis may be due to subjective interpretation by the pathologist. The K-M curves for tumour grade 1 and grade3 are widely separated, it suggests a significant difference in survival between the groups. Based on Table 4.9, the median grade-1 is 32 months, and this group of patients has a better survival probability compared to tumour grade-3 with poorer prognosis. That suggests that in tumour grade 3, patients require a more aggressive therapy. The tumour grade provides valuable information, however, other factors like tumour stage, patients characteristic and specific cancer type should be considered when assessing overall prognosis and treatment options.

Sub-Figure 4.3 h and Table 4.10, the T-stages is determined by physical examination, imaging tests, and biopsies. When T-stage is high generally indicates

a more advanced breast cancer with a poorer prognosis. According to K-M curves, patients who were diagnosed with T4 stage tumour experienced death the most as compared to T1, T2, and T3. However, K-M curves are crossing each other approximately the first 10 months which suggests that they are clinically irrelevant at that stage. Table 4.10, reveals that Stage T4 category patients, their median survival time was 16 months, and 67 patients died at the end study out of the 78 patients. The tumour of these patients (stage T4) had also spread to distant organs like the lungs and brain.

Sub-figure 4.3 i, and Table 4.11, The N-stage of breast cancer patients as defined by AJCC, it is a crucial component of the overall TNM staging system, which is used to determine the extent to which breast cancer cells have spread and, consequently, help and guide the Physicians treatment decision. This is based on information obtained from surgery, including physical examination, imaging tests, and biopsies. According to Sub-Figure 4.3 i, breast cancer patients with stage N0 their survival rate is the highest, understandably in this stage, no regional lymph nodes are involved. Stages N1 and N2 the K-M curves are crossing each other at different times which suggests that it is not statistically meaningful or clinically relevant. The BC patients in stage N3 show lowest survival probability compared to stage NO group, consequently, higher risk of death. Table 4.11 shows that the median survival time of N0-stage group is at least 39 months, meaning that half of BC patients in this category will experience death beyond 39 months. The median survival time of stage N3 group is 16 months, means that half of the people with that stage are expected to live for at least 16 months after being diagnosed.

Sub-figure 4.3 j and Table 4.12, the M-stage is determined using imaging tests such as X-rays or CT scans. A higher M-stage (M1-Stage) indicates a cancer that has spread to other distant organs e.g liver, lungs and brain with a poorer prognosis. In our study, K-M curves show that M1-stage breast cancer patients experienced worse deaths as compared to M0, confirming our understanding that the M1-stage group has experienced distant metastasis. Table 4.12 the median survival time in this

group (M1-stage) was 16 months, means 50% of BC patients with this type of cancer are expected to live at least 16 months after being diagnosed.

Sub-figure 4.3 k and Table 4.13, the breast cancer patients according to AJCC group stage 1, where the tumour size is very small, often less than 2cm in diameter and cancer has not spread to nearby lymph nodes or distant organs such as lungs, liver or bones, the mortality rate is extremely low. Patients according to AJCC group 4 definition, the cancer cells have now spread to distant organs such as the liver, lungs or brain, here the death rate of breast patients is expected to be very high. In AJCC stage 2 and 3, the Kaplan-Meier curves cross each other at certain points. This suggests that there is no much observed difference between these 2 groups (AJCC stage 2 and 3). A statistical test (e.g. log-rank test) would be needed to determine if this crossing is statistically significant, and it represents a true difference between the groups. The AJCC stage 4, the K-M curves show that the survival probability is smaller, which means higher risk of death. In Table 4.13, the AJCC stage group 4 patients is 15 months less than 30 months of group 2 patients which suggests a worse overall survival death. Thus, 50% of BC patients with this stage are expected to live at least 15 months after being diagnosed.

Sub-figure 4.3 i and Table 4.14. The surgery is typically recommended by surgeons for invasive breast cancer as this type of cancer can spread to other parts of the body such as lungs, liver or brain. The lumpectomy or mastectomy is considered by doctors if the tumour is smaller and has not spread to lymph nodes or the tumour is large and spread to lymph nodes, respectively. Our analysis reveals (K-M curves by surgery) that the breast cancer patients who underwent surgery survived longer compared to those who did not. Surgery may be followed by additional treatments, such as chemotherapy, radiation therapy, or hormone therapy, to reduce the risk of recurrence. Table 4.14 shows that the median survival time of breast patients that were operated is 36 months compared to 15 months without surgery. The BC patients who underwent surgery 50% survived beyond 36 months.

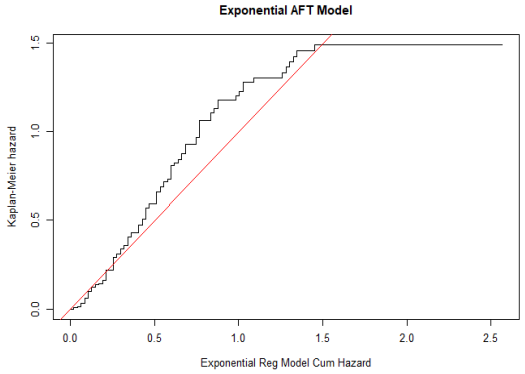
Sub-Figure 4.3 m and Table 4.15. The K-M curves cross each other after 24 months suggesting that these groups experience different chemotherapy treatment effects. The group that was treated with chemotherapy had initially a higher survival rate, thereafter the Kaplan-Meier curves kept crossing each other which suggests that the crossing is not statistically meaningful or clinically relevant. Furthermore, in Table 4.15, the median of cancer patients that were treated with chemotherapy and without were basically the same, which means that chemotherapy treatment was overall no longer effective after 26 months.

Sub-Figure 4.3 n and Table 4.16, the breast cancer patients that were treated with endocrine therapy had higher survival rates than patients without endocrine therapy treatment. That suggests that endocrine therapy is very effective in treating breast cancer patients. Based on Table 4.16 the median of breast cancer patients who were treated with endocrine therapy was 35 months compared with 19 months without which suggests a better overall survival and a lower risk of death.

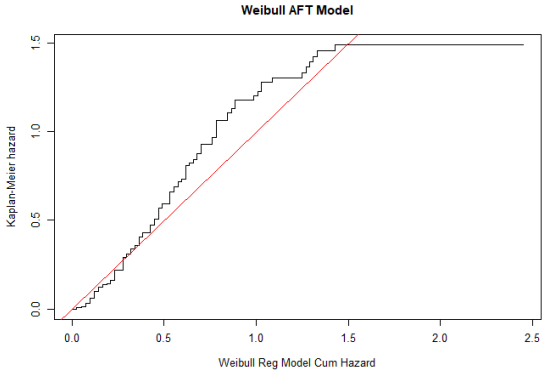
4.3 Models diagnostics

Figure 4.1 shows the Cox-Snell residuals plots fitting the exponential, Weibull, lo-logistic, log-normal and gamma models with covariates age at baseline, menopause, HER2, CK5/6, Subtype triple negative, Subtype Luminal, Subtype Triple negative (basal), Tumour Grade, Tumour stage T, Stage N, Stage M, AJCC Stages, Surgery, Chemotherapy, Endocrine therapy, We fitted these covariates in parametric AFT models in Figure 4.1 using R package version 4.3.1. The exponential, Weibull and Gamma models were not fitting well as they deviate from the straight line, that is, Sub-figure 4.1(a), Sub-figure 4.1(b) and Sub-figure 4.1(e). However, log-logistic and log-normal look similar and are close to a straight line with unit slope and zero intercept. There is no significance difference between these two graphs. The results are similar to those obtained in Table 4.1 (Akaike Information Criterion of six distributions fitted to the full model).

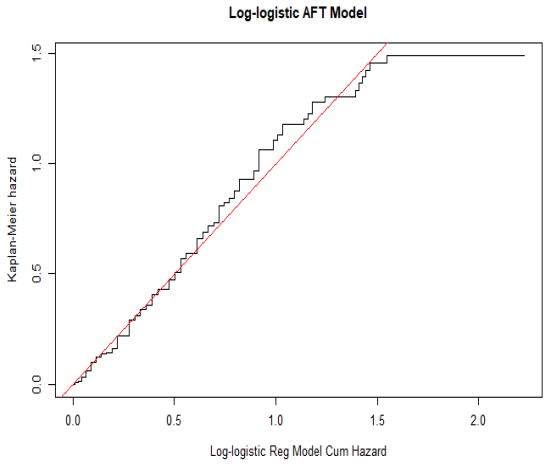
The difference among these two distributions are small with log-normal distribution having a smaller AIC value which could have been influenced by one or two parameters compared with Log-logistic distribution. However, the log-logistic provides the most used AFT model for such situations. The log-logistic distribution is a non-monotonic hazard function that increases at early times and decreases at later stages. It is similar in shape to log-normal distribution, but its cumulative distribution function has a simple closed form, which becomes important computationally when fitting data with censoring (Clayton & Cuzick, 1985). Hence, Log-logistic will be our final preferred model for our right censored distribution.



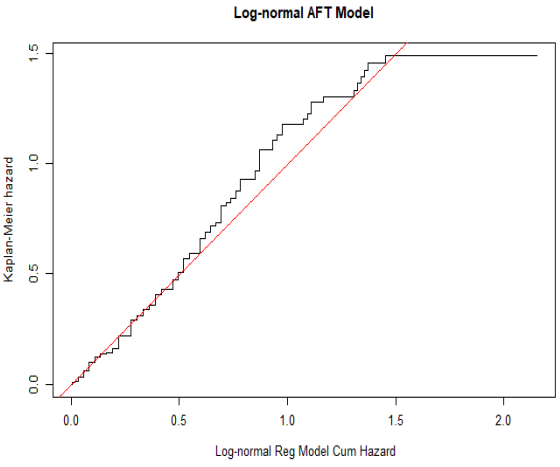
a) Exponential AFT Model



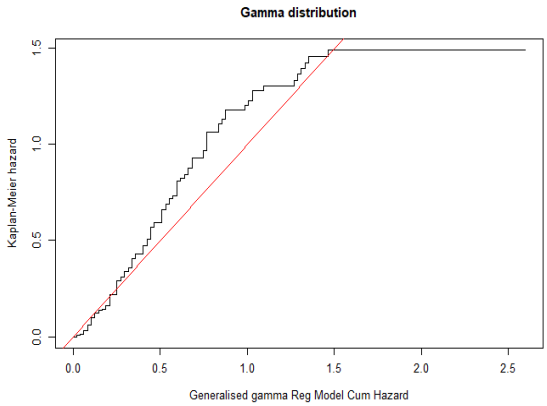
b) Weibull AFT model



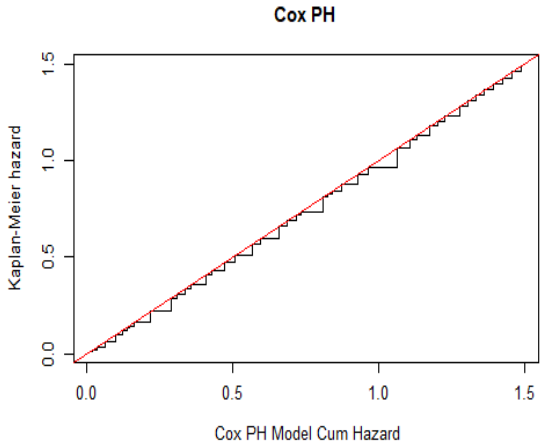
c) Log-logistic AFT model



d) log-normal AFT model



e) Gamma AFT Model



f) Cox PH model

Figure 4. 2: Cox-Snell residuals plot for different accelerated failure time models

Table 4. 15: Akaike information criterion of six distributions fitted to the full model.

Model	Log-likelihood	Number of covariates	Number of parameters	AIC
Exponential	-532.983	30	31	1067.966
Weibull	-532.745	30	31	1069.49
Log-logistic	-522.0769	30	31	1048.154
Log-normal	-521.9172	30	31	1047.834
General Gamma	-532.9647	30	31	1069.929
Cox PH	-482.6188	26	27	414.6035

4.3.1 Results and discussion with log-logistic model

Table 4.16: Hazard ratio from Log-logistic proportional hazard model for Breast Cancer patient's dataset in Pietersburg Hospital in Limpopo Province.

Univariate Analysis					
	Parameters estimate ($\boldsymbol{\beta}_i$)	std. error	HR (exp ($\boldsymbol{\beta}_i$))	95% CI	P-value
(Intercept)	4.84	1.06	127.	(15.96;1012.32)	0.00000471
Age_	-0.0117	0.00962	0.988	(0.97;1.01)	0.223
Premenopause	-0.170	0.283	0.844	(0.48;1.47)	0.548
Postmenopause (Ref)	0.000	1.000			
OR_Positive	1.23	0.921	3.42	(0.56;20.7)	0.182
OR_Negative (Ref)	0.000	1.000			
PR_Positive	-0.156	0.216	0.855	(0.56;1.31)	0.469
PR_Negative (Ref)	0.000	1.000			
HER2_Positive	-1.10	0.736	0.332	(0.08;1.4)	0.134
HER2_Negative (Ref)	0.000	1.000			
CK56_Positive	0.102	0.248	1.11	(0.68;1.8)	0.68
CK56_Negative (Ref)	0.000	1.000			
Subtype_HER2-like	1.41	0.667	4.08	(1.1;15.03)	0.0349
Subtype_Luminal-A	-1.06	0.804	0.346	(0.07;1.67)	0.186
Subtype_Luminal-B	0.000	1.000			
Grade-1	0.498	0.366	1.64	(0.8;3.35)	0.174
Grade-3	-0.237	0.188	0.789	(0.55;1.14)	0.207
Grade-2	0.000	1.000			
T_stage_T2	-0.407	0.509	0.666	(0.25;1.81)	0.424

T_stage_T3	-0.480	0.570	0.619	(0.2;1.89)	0.4
T_stage_T4	-1.16	0.579	0.314	(0.1;0.98)	0.0451
T_stage_T1 (Ref)	0.000	1.000			
N_stage_N1	-0.0929	0.251	0.911	(0.56;1.49)	0.711
N_stage_N2	0.0565	0.303	1.06	(0.58;1.92)	0.852
N_stage_N3	0.526	0.401	1.69	(0.77;3.71)	0.189
N_stage_N0 (Ref)	0.000	1.000			
M_stage_M1	0.345	0.642	1.41	(0.4;4.95)	0.591
M_stag_M0 (Ref)	0.000	1.000			
AJCC_stage_Stage-2	-0.932	0.541	0.394	(0.14;1.14)	0.0851
AJCC_stage_Stage-3	-0.262	0.613	0.770	(0.23;2.56)	0.67
AJCC_stage_Stage-4	-1.55	0.881	0.211	(0.04;1.19)	0.0777
AJCC_Stage-1 (Ref)	0.000	1.000			
Surgery_Yes	0.235	0.216	1.27	(0.83;1.93)	0.275
Surgery_No (Ref)	0.000	1.000			
Chemotherapy_Yes	0.430	0.445	1.54	(0.64;3.67)	0.334
Chemotherapy_No (Ref)	0.000	1.000			
ET_Yes	0.334	0.216	1.40	(0.91;2.13)	0.122
ET_No (Ref)	0.000	1.000			
Log(scale)	-1.06	0.112	0.348	(0.28;0.43)	3.62E-21
Scale	0.348	0.348	NA	(0.72;2.8)	NA

4.3.2 Results and discussion with Cox Proportional Hazard model

Table 4. 4: Hazard ratio from Cox proportional hazard model for Breast Cancer patient's dataset in Pietersburg Hospital in Limpopo Province.

Univariate Analysis					
	Parameters estimate ($\boldsymbol{\beta}_i$)	Std error	HR (exp($\boldsymbol{\beta}_i$))	95%CI	P-value
Age_	0.0293	0.0192	0.971	(0.99;1.07)	0.128
Premenopause	0.258	0.567	0.772	(0.43;3.94)	0.649
Postmenopause (Ref)	0.000	1.000			
OR_Positive	-1.93	1.94	6.89	(0;6.42)	0.319
OR_Negative (Ref)	0.000	1.000			
PR_Positive	0.0789	0.473	0.924	(0.43;2.75)	0.868
PR_Negative (Ref)	0.000	1.000			
HER2_Positive	2.22	1.53	0.108	(0.46;183.09)	0.145
HER2_Negative (Ref)	0.000	1.000			
CK5/6_Positive	-0.290	0.492	1.34	(0.29;1.96)	0.556
CK56_Negative (Ref)	0.000	1.000			
Subtype_HER2-like	-3.13	1.41	22.8	(0;0.69)	0.0262
Subtype_Luminal-A	1.56	1.68	0.211	(0.18;129.02)	0.355
Grade_Grade-1	-0.658	0.731	1.93	(0.12;2.17)	0.368
Grade_Grade-3	0.626	0.407	0.535	(0.84;4.14)	0.124
Grade-2	0.000	1.000			
T_stage_T2	0.493	1.06	0.611	(0.2;13.2)	0.643
T_stage_T3	0.733	1.22	0.480	(0.19;22.87)	0.548
T_stage_T4	1.69	1.19	0.185	(0.52;55.7)	0.157

T_stage_T1 (Ref)	0.000	1.000			
N_stage_N1	0.114	0.520	0.892	(0.4;3.1)	0.826
N_stage_N2	-0.206	0.588	1.23	(0.26;2.58)	0.726
N_stage_N3	-1.49	0.889	4.46	(0.04;1.28)	0.0929
N_stage_N0 (Ref)	0.000	1.000			
M_stage_M1	-1.33	1.36	3.78	(0.02;3.82)	0.329
M_stag_M0 (Ref)	0.000	1.000			
AJCC_stage_Stage-2	1.80	1.28	0.165	(0.5;74.44)	0.158
AJCC_stage_Stage-3	0.698	1.48	0.497	(0.11;36.97)	0.638
AJCC_stage_Stage-4	4.17	2.04	0.0155	(1.18;3533.34)	0.0412
AJCC_Stage-1 (Ref)	0.000	1.000			
Surgery_Yes	-1.00	0.464	2.72	(0.15;0.91)	0.0309
Surgery_No (Ref)	0.000	1.000			
Chemotherapy_Yes	-0.535	0.980	1.71	(0.09;4.01)	0.586
Chemotherapy_No (Ref)	0.000	1.000			
ET_Yes	-0.554	0.430	1.74	(0.25;N/A)	0.198
ET_No (Ref)	0.000	1.000			

4.3.2.1 Menopause

The premenopausal breast cancer women had HR=0.772, meaning an individuals with breast cancer have 22.8% lower risk of experiencing death compared to postmenopause women. Since CI (0.43;3.94) and p-value= 0.649, menopause as covariates not statistically significant, however, it might have meaningful impact on

patient outcome. The standard error equals 0.567 indicating coefficient estimate of premenopausal was not precise.

4.3.2.2 Positive oestrogen receptor

Breast cancer women with positive oestrogen receptor (OR-positive) who received chemotherapy or endocrine therapy had HR =0.281. That suggest that individuals with a positive oestrogen receptor status have a lower risk of death compared to negative oestrogen receptor. Consequently, the risk of death is 0.281 times lower for individuals with positive oestrogen receptor compared to negative oestrogen receptor. This is a big effect size, indicating a weak association between oestrogen receptor and death. The C.I (1.00; 614.00) and this interval includes 1 and p-value= 0.319 indicates that it is not statistically significant.

4.3.2.3 Progesterone receptor

The BC patients with positive progesterone receptor have HR=1.082, indicating that individual with positive progesterone receptor have approximately 8.2% higher risk of death compared to those with negative progesterone receptor. Secondly, these women with positive progesterone receptor are more likely to experience a recurrence compared with negative progesterone receptor. The C.I (0.43;2.75) and p-value=0.868, suggest that positive progesterone receptor is not statistically significant. The standard error equals 0.473 indicating that the estimation of positive progesterone receptor parameter was not precise.

4.3.2.4 Human epidermal growth receptor factor 2

Women with positive human epidermal growth receptor factor 2 (HER2 positive) have HR=0.108, meaning that individuals with positive HER2 have approximately 89.2% lower risk of death compared with HER2 negative. Since HR=0.108 which is small it might not have a substantial impact on patient’s death in practice, especially in the context of surgery, chemotherapy and endocrine therapy treatment. The p-value=0.145 means that HER2 positive is not statistically significant, and standard

error is 1.53 shows a relatively large degree of uncertainty in the estimation of HER2 positive parameter.

4.3.2.5 Cytokeratin 5/6.

The positive cytokeratin 5/6 (CK5/6) BC women had HR=0.748, which means that individuals with a positive CK5/6 test is 0.748 less likely to experience death compared to CK 5/6 negative test. That is, breast cancer patients with a positive CK5/6 tests have a lower risk of death compared to negative CK5/6 tests. The p-value = 0.556, we conclude that CK5/6 positive is not statistically significant. The standard error equals 0.492 which is relatively large that indicate uncertainty in the estimation of CK5/6 positive parameter.

4.3.2.6 Luminal Subtype

A HR=4.759 of luminal subtype A, and non-luminal subtype like HER2 triple negative (basal) indicate that's individuals breast cancer patients with luminal subtype A are approximately 4.759 times more likely to experiencing death compared to luminal subtype B. The effects of luminal subtype A in different subgroups (e.g., age, tumour stage)will help identify patients who may benefit most from specific treatments. The p-value=0.355 means that it is not statistically significant.

4.3.2.7 The tumour grade 2 and grade 3.

Breast cancer women with tumour grade 2 are approximately 1.93 times more likely to experience death compared to tumour grade 1. That is, BC women with tumour grade 2 have a significant higher risk of death compared to those with tumour grade 1. However, HR=1.93 which is high, might not be clinically meaningful it if it does not translate to a substantially difference in patient's death. The p-value=0.368 indicating that tumour grade 2 is not statistically significant

The BC women with tumour grade 3 have HR=0.535, indicating that individuals with tumour grade 3 are approximately 47% less likely to explain death compared to tumour grave 1.

The possible explanation for tumour grade 3 be less than tumour grade 2 would be.

1) Grade 3 tumour might be more aggressive and detectable at earlier stage and leading to earlier diagnosis and treatment.
2) Other factors like age, overall health might had influence tumour grade 3 and death.

4.3.2.8 Tumour T stages

The BC women with tumour T stage 2, T stage 3 and T stage 4 have hazard ratios 1.637, 2.081 and 5.419, respectively, compared with tumours stage T 1 BC patients. A HR=1.637 for T2 tumours means that individuals with T2 tumours are 1.637 more likely to die in any given time compared with T 1 tumours. Similarly for T3 and T4 tumours BC patients. Thus, the tumour T stages advanced from T2 to T4 with a risk of death increasing significantly. The clinical implications of these hazard ratios can help to predict the prognosis of BC patients with different tumour stages, and inform treatment decisions such as choice of treatment modality. Secondly, it can help classify BC patients into risk groups for clinical trials. Their p-values are greater than 0.05 indicating that they are not statistically significant. The standard deviation is relatively big meaning that parameter estimates were not precise.

4.3.2.9 Tumour N stages

A HR=1.121 stage N1 BC women indicate that the death risk is 1.21 times compared to stage N0.That is, individuals in the group with HR=12.1 slightly higher of expressing death compared to those in stage N0. The p-value=0.826, indicating that it is not statistical significant. The hazard ratios of stage N2 and N3 are 1.228 and 4.459 respectively, and these hazard ratios indicate clearly the increasing trend risk with increasing lymph nodes involvement. The stage N2 have a 1.23 times higher risk of death compared to those with N0 stage, and stage N3 have a 4.46 times higher risk of death compared to those with stage N0. Their p-values are greater than 0.05, hence they are not statistically significant.

4.3.2.10 M-stage

The BC woman with HR=3.78 that individuals with M1 stage cancer are 3.78 times more likely to die compared to those with M0 stage cancer. The p-value=0.329. We can conclude that the observed difference in risk is due to chance. This large HR =3.378 have important implications for treatment decisions and patient counselling.

4.3.2.11 AJCC stages.

The BC woman with a HR=0.165, means AJCC with stage 2 cancer have a lower risk of death compared to AJCC stage 1. BC factors like age, and comorbidities may also had an impact on risk of death, thus, AJCC stage 2 death rate seem to less than AJCC stage 1.

A HR=2.010 means that individuals with this stage cancer is 101% more likely to die compared to AJCC stage 1. The HR=3.22 means patients with AJCC stage 4 cancer are 222% more likely to die compared to AJCC stage 1. However, their p-values are greater than 0.05 meaning they are not statistically significant

4.3.2.12 Surgery

A HR=2.72 for surgery indicates that individuals who underwent surgery have a 2.72 times higher risk cancer recurrence compared to individuals who did not undergo surgery. Secondly, 172% more likely to experience death compared to AJCC stage 1. We can confidently, conclude that it is statistically significant or the observed risk is not due to chance.

4.3.2.13 Chemotherapy

A HR=1.707 for chemotherapy, indicates that individuals who received chemotherapy have a 1.71 times higher risk of death compared to individuals who did not receive chemotherapy. The hazard ratio equals 1.707 means it is not statistical significance because p-value=0.586, we can conclude confidently that the observed difference was due to chance, an HR=1.71 suggests a great difference

risk, which may have important clinical implications for treatment decision and counselling. The timing of administering chemotherapy treatment can also influence hazard ratio.

4.3.2.14 Endocrine Therapy.

A HR=1.74 for endocrine therapy, indicates that individuals who received endocrine therapy have a 1.74 times higher risk of death compared to individuals who did not receive endocrine therapy. The hazard ratio equals 1.74 means it is not statistical significant because p-value=1.98, we can conclude confidently that the observed difference was due to chance, an HR=1.74 suggests a great difference risk, which may have important clinical implications for treatment decision and counselling. The timing of administering endocrine treatment can also influence hazard ratio.

4.4 Chapter summary

Chapter 4 presented and analysed the results pertaining to prognostic factors of survival time for breast cancer patients in Petersburg Oncology Hospital in Limpopo Province, South Africa. In the analysis, Kaplan Meier curve method, log-logistic and Cox Proportional hazard models were employed to analyse the following factors, namely, menopause status, oestrogen receptors, progesterone receptors, Cytokeratin 5/6, luminal subtype, tumour grades, human epidermal growth factor 2, T stages, N stages, AJCC stages, chemotherapy treatment, endocrine therapy and whether breast cancer women underwent surgery or not. The fitted models (Cox PH, log-logistic) and K-M method were able to identify those factors that contributed to better or poorer prognosis of breast cancer women in Pietersburg Hospital.

Chapter 5

Conclusion and Recommendation

5.1 Introduction

This chapter covers the conclusions, recommendations and the additional research topics. The aim of this study was to identify prognostic factors that contribute to breast cancer in Pietersburg Oncology Hospital by comparing proportional hazard and accelerated failure time models, assessing the performance of prognostic and predictive models, and stratifying patients into risk categories.

5.2 Conclusion

When using Table 4.17(AIC), the Cox proportional hazard function appeared to be better compared to the log-logistic model. The appropriateness of the proportional hazard assumptions was checked using log (-log) plots. Since our primary interest was to identify factors that influence the relative risk of death, and Cox model seemed to be the better candidate because is able to handle temporal effects of the covariates (Lu *et. al.,*2005). For example, the chemotherapy worked well at the initial stage Figure 4.3 m, treatment, but gradually lost its potency due to cancer cell metastasis.

It was found that premenopausal breast cancer patients lived longer than postmenopausal patients. The difference in survival time between premenopausal and postmenopausal breast cancer patients highlights the importance of understanding how hormonal, biological, and demographic factors influence cancer

prognosis. This finding has implications for personalised treatment, and healthcare policy aimed at improving survival for all post-menopause breast cancer patients. That is confirmed by previous studies done by Karla *et al.,* (2010), Marzeena *et al.,* (2015). The oestrogen receptor (OR positive) status also played an important role for the BC patients' survival. Patients with OR-positive breast cancer had lived longer on overall. The difference in survival between OR-positive and OR-negative breast cancer patients highlights how OR status serves as a key factor in prognosis and treatment planning. The OR-positive patients benefit from targeted hormone therapies like endocrine therapy, which contribute to longer survival. However, OR-negative patients have an aggressive nature of the disease and limited treatment options result in shorter survival. The study by Metcalfe *et al.,* (2019), Newman *et al.,* (1997), Hammond *et al.,* (2010) agree with our findings.

The breast cancer patients with human epidermal growth factor 2 (HER2) positive survived longer in Pietersburg Oncology Hospital due to availability of endocrine therapy. The HER2 is very important and critical to be considered for further survival improvement across all breast cancer suffers previous studies have reported similar finding. The study done by Fei *et al.,* (2021), Yoon *et al.,* (2022) and Wolff et al., (2006) agree that HER is vital and play a key role for survival improvement.

The breast cancer patients in AJCC stage 4 and M1 stage died the most due to advanced cancer cells that had spread to distant organs. That was as a result of late screening of BC patients as supported by Taskindoust *et al.,* (2021), Soerjomataram *et al.,* (2008) and Soerjomataram *et al.,* (2008).

The breast cancer women who were in category stage N3 and T4 tumours were diagnosed at later stages experienced the most deaths compared to other earlier stages. The rationale is that tumours at this stages are more aggressive and they grow and spread rapidly. Our finding were confirmed by Saadatmand *et al.,* (2015) and Berndt *et al.,* (1969).

Patients who underwent surgery found to have a higher risk of recurrence and death in Pietersburg Oncology Hospital. Our finding agrees with study done by Kingsmore *et al.,* (2004).

5.3 Recommendation

The biopsies will offer the potential for non-invasive, real-time monitoring of disease progression, as well as recurrence and resistance to treatment like chemotherapy. Recent studies have suggested that neutrophilia may be a prognostic factor for breast cancer. Future research is needed for ER-negative to improve treatment effects options that will result in this group living longer. Further research is needed to understand the underlying mechanisms

5.4 Limitation of the study

Our sample of 153 patients may limit the generalisation of our findings to the broader population and the smaller sample size may affect the statistical power of our study.

The study used secondary data and some important information was not captured correctly which resulted in a reduced sample size. In the beginning, we had 303 patients but we ended up with a reduced sample size of 153.

We were not able to conduct randomised controlled trials (RCTs) for breast cancer research by assigning patients to different treatment groups due to the use of secondary data. Furthermore, our study relied on retrospective data, which may be subject to biases and limitations in data collection and reporting.

Individual variations in breast cancer susceptibility and response to treatment could have complicated our research findings. Breast cancer is a highly complex disease with various subtypes, each having unique characteristics and responses to treatment. This heterogeneity makes it challenging to develop universal approaches to prevention, screening, and treatment.

The AJCC staging system is periodically updated to reflect advances in diagnostic techniques and understanding of cancer biology. Our study might have utilised an older version of the AJCC staging and therefore, there may be a need to interpret the results cautiously as the definition of stages may have changed. Molecular markers such as ER, PR and HER2 can give additional prognostic information beyond the AJCC stage.

5.5 Recommendation for Future Research

The biopsies will offer the potential for non-invasive, real-time monitoring of disease progression, as well as recurrence and resistance to treatment. By using ctDNA to detect minimal residual disease, clinicians may need to identify patients at higher risk of relapse. Secondly, analysing ctDNA can help identify genetic alterations that may be driving tumour growth and resistance to therapy, which could lead to more personalised treatment options. Understanding the role of tumour-associated macrophages (TAMs) in promoting tumour growth and resistance to therapy could lead to new therapeutic strategies. Recent studies have suggested that neutrophilia may be a prognostic factor for breast cancer. Further research is needed to understand the underlying mechanisms.

References

Aalen, O., 1978. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pp.701-726.

Abi Jaoude, J., Kayali, M., de Azambuja, E., Makki, M., Tamim, H., Tfayli, A., El Saghir, N., Geara, F., Piccart, M., Poortmans, P. and Zeidan, Y.H., 2020. De-intensifying radiation therapy in HER-2 positive breast cancer: to boost or not to boost?. International Journal of Radiation Oncology* Biology* Physics, 108(4), pp.1040-1046.

Alanko A, Heinonen E, Scheinin TM, et al., Oestrogen and progesterone receptors and disease-free interval in primary breast cancer. Br J Cancer 1984;50:

Alif, B.U., 2022. Cancer prognosis & Survival rate prediction using machine learning algorithm (Doctoral dissertation, Department of Electronic and Telecommunication Engineering).

Almstedt, K., Heimes, A.S., Kappenberg, F., Battista, M.J., Lehr, H.A., Krajnak, S., Lebrecht, A., Gehrmann, M., Stewen, K., Brenner, W. and Weikel, W., 2022. Long-term prognostic significance of HER2-low and HER2-zero in node-negative breast cancer. European Journal of Cancer, 173, pp.10-19.

An, D., Choi, J., Lee, J., Kim, J.Y., Kwon, S., Kim, J., Lee, S., Jeon, S., Lee, C., Lee, S. and Woo, H., 2022. Time to surgery and survival in breast cancer. BMC surgery, 22(1), pp.388.

Andersen, P.K. (1982). Testing goodness of fit of Cox's regression and life model survival data (with discussion). Biometrics 38, pp 67-77.

Anyigba, C.A., Awandare, G.A. and Paemka, L., 2021. Breast cancer in sub-Saharan Africa: The current state and uncertain future. *Experimental Biology and Medicine*, *246*(12), pp.1377-1387.

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J.R., Cardoso, F., Siesling, S. and Soerjomataram, I., 2022. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast, 66, pp.15-23.

Ayeni, O.A., Joffe, M., Mapanga, W., Chen, W.C., O'Neil, D.S., Phakathi, B., Nietz, S., Buccimazza, I., Čačala, S., Stopforth, L.W. and Jacobson, J.S., 2023. Multimorbidity and overall survival among women with breast cancer: results from the South African Breast Cancer and HIV Outcomes Study. *Breast Cancer Research*, *25*(1), pp.1-13.

Ayeni, O.A., Joffe, M., Mapanga, W., Chen, W.C., O'Neil, D.S., Phakathi, B., Nietz, S., Buccimazza, I., Čačala, S., Stopforth, L.W. and Jacobson, J.S., 2023. Multimorbidity and overall survival among women with breast cancer: results from the South African Breast Cancer and HIV Outcomes Study. Breast Cancer Research, 25(1), pp.1-13.

Ayeni, O.A., Norris, S.A., Joffe, M., Cubasch, H., Nietz, S., Buccimazza, I., Singh, U., Čačala, S., Stopforth, L., Chen, W.C. and McCormack, V.A., 2020. The multimorbidity profile of South African women newly diagnosed with breast cancer. *International journal of cancer*, *147*(2), pp.361-374.

Ayeni, O.A., Norris, S.A., Joffe, M., Cubasch, H., Nietz, S., Buccimazza, I., Singh, U., Čačala, S., Stopforth, L., Chen, W.C. and McCormack, V.A., 2020. The multimorbidity profile of South African women newly diagnosed with breast cancer. *International journal of cancer*, *147*(2), pp.361-374.

Ayeni, O.A., O'Neil, D.S., Pumpalova, Y.S., Chen, W.C., Nietz, S., Phakathi, B., Buccimazza, I., Čačala, S., Stopforth, L.W., Farrow, H.A. and Mapanga, W., 2022. Impact of HIV infection on survival among women with stage I-III breast

cancer: Results from the South African breast cancer and HIV outcomes study. *International journal of cancer*, *151*(2), pp.209-221.

Azam, S., Eriksson, M., Sjölander, A., Gabrielson, M., Hellgren, R., Czene, K. and Hall, P., 2021. Mammographic microcalcifications and risk of breast cancer. British journal of cancer, 125(5), pp.759-765.

Balhi, S., 2023. Factors associated with delayed diagnosis among sub-Saharan African women. *Indian Journal of Community and Family Medicine*, *9*(1), pp.14.

Baria, E., Morselli, S., Anand, S., Fantechi, R., Nesi, G., Gacci, M., Carini, M., Serni, S., Cicchi, R. and Pavone, F.S., 2019. Label-free grading and staging of urothelial carcinoma through multimodal fibre-probe spectroscopy. Journal of Biophotonics, 12(11), pp.e201900087.

Basha, L. and Hoxha, F., 2019. Kernel estimation of the baseline function in the Cox model. European Scientific Journal, *15*(6), pp.105-118.

Benefield, H.C., Allott, E.H., Reeder-Hayes, K.E., Perou, C.M., Carey, L.A., Geradts, J., Sun, X., Calhoun, B.C. and Troester, M.A., 2020. Borderline estrogen receptor–positive breast cancers in Black and White Women. JNCI: Journal of the National Cancer Institute, 112(7), pp.728-736.

Benson, J.R. and Jatoi, I., 2012. The global breast cancer burden. Future oncology, 8(6), pp.697-702.

Berndt, H. and Titze, U., 1969. TNM clinical stage classification of breast cancer. *International Journal of Cancer*, *4*(6), pp.837-844.

Bhuiyan, M.M.Z.U., Maele, M.M., Mavhungu, R. and Ooko, F., 2022. Breast cancer: Factors influencing late-stage presentation at the Mankweng Hospital breast cancer clinic, Polokwane, Limpopo Province, South Africa. *SAMJ: South African Medical Journal*, *112*(11b), pp.906-910.

Bhuiyan, M.M.Z.U., Maele, M.M., Mavhungu, R. and Ooko, F., 2022. Breast cancer: Factors influencing late-stage presentation at the Mankweng Hospital breast

cancer clinic, Polokwane, Limpopo Province, South Africa. SAMJ: South African Medical Journal, 112(11b), pp.906-910.

Bissell, M.C., Kerlikowske, K., Sprague, B.L., Tice, J.A., Gard, C.C., Tossas, K.Y., Rauscher, G.H., Trentham-Dietz, A., Henderson, L.M., Onega, T. and Keegan, T.H., 2020. Breast cancer population attributable risk proportions associated with body mass index and breast density by race/ethnicity and menopausal status. Cancer Epidemiology, Biomarkers & Prevention, 29(10), pp.2048-2056.

Boyages, J., 2017. Radiation therapy and early breast cancer: current controversies. *The Medical Journal of Australia*, *207*(5), pp.216-222.

Brandão, M., Guisseve, A., Bata, G., Firmino-Machado, J., Alberto, M., Ferro, J., Garcia, C., Zaqueu, C., Jamisse, A., Lorenzoni, C. and Piccart-Gebhart, M., 2021. Survival impact and cost-effectiveness of a multidisciplinary tumor board for breast cancer in Mozambique, sub-Saharan Africa. The oncologist, 26(6), pp.e996-e1008.

Breslow, N.E. (1970) Generalized Kruskall-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 57, pp 579-595.

Breslow, N.E. (1974). Covariance analysis of censored survival data. Biometrics 30, pp 89-99.

Burguin, A., Diorio, C. and Durocher, F., 2021. Breast cancer treatments: updates and new challenges. Journal of personalized medicine, 11(8), pp.808.

Chen, W., Hong, Z., Kang, S., Lv, X. and Song, C., 2022. Analysis of Stemness and Prognosis of Subtypes in Breast Cancer Using the Transcriptome Sequencing Data. Journal of Oncology, 2022.

Chira, C., Kirova, Y.M., Liem, X., Campana, F., Peurien, D., Amessis, M., Fournier-Bidoz, N., Pierga, J.Y., Dendale, R., Bey, P. and Fourquet, A., 2013. Helical tomotherapy for inoperable breast cancer: a new promising tool. *BioMed Research International*, *2013*.

Chlebowski, R.T., Luo, J., Anderson, G.L., Barrington, W., Reding, K., Simon, M.S., Manson, J.E., Rohan, T.E., Wactawski-Wende, J., Lane, D. and Strickler, H.,

2019. Weight loss and breast cancer incidence in postmenopausal women. Cancer, 125(2), pp.205-212.

Christakoudi, S., Tsilidis, K.K., Dossus, L., Rinaldi, S., Weiderpass, E., Antoniussen, C.S., Dahm, C.C., Tjønneland, A., Mellemkjær, L., Katzke, V. and Kaaks, R., 2023. A body shape index (ABSI) is associated inversely with post-menopausal progesterone-receptor-negative breast cancer risk in a large European cohort. BMC cancer, 23(1), pp.1-12.

Chung, S.R., Choi, W.J., Cha, J.H., Kim, H.H., Shin, H.J., Chae, E.Y. and Yoon, G.Y., 2019. Prognostic factors predicting recurrence in invasive breast cancer: An analysis of radiological and clinicopathological factors. Asian journal of surgery, 42(5), pp.613-620.

Citron, M.L., Berry, D.A., Cirrincione, C., Hudis, C., Winer, E.P., Gradishar, W.J., Davidson, N.E., Martino, S., Livingston, R., Ingle, J.N. and Perez, E.A., 2003. Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: first report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. *Journal of clinical oncology*, *21*(8), pp.1431-1439.

Clayton D, Cuzick J. Multivariate generalisations of the proportional hazard model. JR Stat Soc.(1985) 7:pp 82-117

Collett D. 2003. Modeling Survival Data in Medical Research. 2nd Edition, Chapman & Hall/CRS A CRC Press Company, London, New York, Washington, D.C.

Collins, P.M., Brennan, M.J., Elliott, J.A., Abd Elwahab, S., Barry, K., Sweeney, K., Malone, C., Lowery, A., Mclaughlin, R. and Kerin, M.J., 2021. Neoadjuvant chemotherapy for luminal a breast cancer: Factors predictive of histopathologic response and oncologic outcome. The American Journal of Surgery, 222(2), pp.368-376.

Cortesi, L., Rugo, H.S. and Jackisch, C., 2021. An overview of PARP inhibitors for the treatment of breast cancer. Targeted oncology, 16(3), pp.255-282.

Cox, D.R. and Snell, E.J., (1968). A general definition of residuals. Journal of the Royal Statistical Society: Series B (Methodological), 30(2), pp.248-275.

Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), pp.187-202.

Criscitiello, C., Vingiani, A., Maisonneuve, P., Viale, G., Viale, G. and Curigliano, G., 2020. Tumor-infiltrating lymphocytes (TILs) in ER+/HER2− breast cancer. Breast cancer research and treatment, 183, pp.347-354.

Crowley, J. and Hu, M (1977). Covariance analysis of heart transplant survival data. J. Am. Stat. assoc. 72, pp27-36.

Czajka, M.L. and Pfeifer, C., 2020. Breast cancer surgery.

Dar, H., Johansson, A., Nordenskjöld, A., Perez-Tenorio, G., Yau, C., Benz, C.C., Esserman, L.J., Nordenskjöld, B., Stål, O., Fornander, T. and Lindström, L.S., 2023. Abstract P2-03-04: Long-Term Benefit from Adjuvant Tamoxifen in Luminal A and Luminal B Breast Cancer Patients. Cancer Research, 83(5_Supplement), pp.P2-03.

Davey, M.G., Ryan, É.J., Folan, P.J., O'Halloran, N., Boland, M.R., Barry, M.K., Sweeney, K.J., Malone, C.M., McLaughlin, R.J., Kerin, M.J. and Lowery, A.J., 2021. The impact of progesterone receptor negativity on oncological outcomes in oestrogen-receptor-positive breast cancer. BJS open, 5(3), pp.zrab040.

De Jager, T., 2015. Considering the impact of anonymity on the quality of data obtained in an insight community. In *Annual conference delegate copy*.

Denkert, C., Seither, F., Schneeweiss, A., Link, T., Blohmer, J.U., Just, M., Wimberger, P., Forberger, A., Tesch, H., Jackisch, C. and Schmatloch, S., 2021. Clinical and molecular characteristics of HER2-low-positive breast

cancer: pooled analysis of individual patient data from four prospective, neoadjuvant clinical trials. The Lancet Oncology, 22(8), pp.1151-1161.

Dimitrov, G., Atanasova, M., Popova, Y., Vasileva, K., Milusheva, Y. and Troianova, P., 2022. Molecular and genetic subtyping of breast cancer: the era of precision oncology. WCRJ, 9, pp.e2367.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. J. AM, stat. Assoc. 72, pp 557-565.

Elmore, S.N.C., Mushonga, M., Iyer, H.S., Kanda, C., Chibonda, S., Chipidza, F., Makunike Mutasa, R., Muchuweti, D., Muguti, E.G., Maunganidze, A. and Ndlovu, N., 2021. Breast cancer in Zimbabwe: patterns of care and correlates of adherence in a national referral hospital radiotherapy center cohort from 2014 to 2018. *Cancer Medicine*, *10*(11), pp.3489-3498.

Elmore, S.N.C., Mushonga, M., Iyer, H.S., Kanda, C., Chibonda, S., Chipidza, F., Makunike Mutasa, R., Muchuweti, D., Muguti, E.G., Maunganidze, A. and Ndlovu, N., 2021. Breast cancer in Zimbabwe: patterns of care and correlates of adherence in a national referral hospital radiotherapy center cohort from 2014 to 2018. Cancer Medicine, 10(11), pp.3489-3498.

Erichsen, J., 2023. The Science and Art of Surgery: Vol. I. BoD–Books on Demand.

Faruk, A., 2018, March. The comparison of proportional hazards and accelerated failure time models!in analyzing the first birth interval survival data. In Journal of Physics: Conference Series (Vol. 974, No. 1, p. 012008). IOP Publishing.

Fei, Fei, Gene P. Siegal, and Shi Wei. "Characterization of estrogen receptor-low-positive breast cancer." *Breast cancer research and treatment* 188 (2021): 225-235.

Fernandez-Moya, A., Morales, S., Arancibia, T., Gonzalez-Hormazabal, P., Tapia, J.C., Godoy-Herrera, R., Reyes, J.M., Gomez, F., Waugh, E. and Jara, L., 2020. Germline variants in driver genes of breast cancer and their association

with familial and early-onset breast cancer risk in a Chilean population. *Cancers*, *12*(1), pp.249.

Fina, E., 2022. Signatures of Breast Cancer Progression in the Blood: What Could Be Learned from Circulating Tumor Cell Transcriptomes. Cancers, 14(22), p.5668.

Fleming, TR and Harrington, D.P. (1981). A class of hypothesis tests for one and two samples of censored survival data. Commun. Stat. 10, pp763-.795.

Foerster, M., McKenzie, F., Zietsman, A., Galukande, M., Anele, A., Adisa, C., Parham, G., Pinder, L., Schüz, J., McCormack, V. and dos-Santos-Silva, I., 2021. Dissecting the journey to breast cancer diagnosis in sub-Saharan Africa: Findings from the multicountry ABC-DO cohort study. *International journal of cancer*, *148*(2), pp.340-351.

Fortner, R.T., Sisti, J., Chai, B., Collins, L.C., Rosner, B., Hankinson, S.E., Tamimi, R.M. and Eliassen, A.H., 2019. Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: results from the Nurses' Health Studies. Breast Cancer Research, 21(1), pp.1-9.

Fox,J(2016). Applied Regression Analysis and Generalised linear models.Sage, Thousand Oakes, CA, Third edition.

Francies, F.Z., Hull, R., Khanyile, R. and Dlamini, Z., 2020. Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options.

Freitas AG, Weller M. Patient delays and system delays in breast cancer treatment in developed and developing countries. Cien Saude Colet 2015;20:3177-89.

Gakunga, R., Kinyanjui, A., Ali, Z., Ochieng', E., Gikaara, N., Maluni, F., Wata, D., Kyeng', M., Korir, A. and Subramanian, S., 2019. Identifying barriers and facilitators to breast cancer early detection and subsequent treatment engagement in Kenya: a qualitative approach. *The oncologist*, *24*(12), pp.1549-1556.

Gathirua-Mwangi, W.G., Palmer, J.R., Champion, V., Castro-Webb, N., Stokes, A.C., Adams-Campbell, L., Marley, A.R., Forman, M.R., Rosenberg, L. and Bertrand, K.A., 2022. Maximum and time-dependent body mass index and breast cancer incidence among postmenopausal women in the black women's health study. American journal of epidemiology, 191(4), pp.646-654.

Gebremariam, A., Addissie, A., Worku, A., Assefa, M., Pace, L.E., Kantelhardt, E.J. and Jemal, A., 2019. Time intervals experienced between first symptom recognition and pathologic diagnosis of breast cancer in Addis Ababa, Ethiopia: a cross-sectional study. *BMJ open*, *9*(11), pp.e032228.

Gehan, E.A., 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, *52*(1-2), pp.203-224.

Gehan, E.A.A (1965). A generalized Wilcoxon on test for comparing arbitrarily single censored sample. Biometrika 52,pp 203-223.

Gelber, R.D., Goldhirsch, A., and Cole, B.F.(1993). Parametric extrapolation of survival estimates with application toquality of life evaluation of treatments. Control clinical trails, 14, pp. 485-499

Getachew, S., Tesfaw, A., Kaba, M., Wienke, A., Taylor, L., Kantelhardt, E.J. and Addissie, A., 2020. Perceived barriers to early diagnosis of breast Cancer in south and southwestern Ethiopia: a qualitative study. *BMC women's health*, *20*, pp.1-8.

Gewirtz, D., 1999. A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. *Biochemical pharmacology*, *57*(7), pp.727-741.

Giulianelli, S., Lamb, C.A. and Lanari, C., 2021. Progesterone receptors in normal breast development and breast cancer. Essays in Biochemistry, 65(6), pp.951-969.

Golijanin, D., Radovanovic, Z., Radovanovic, D., Djermanovic, A., Djuric, M., Zahorjanski, S., Lukic, D., Ignjatovic, M.K. and Protic, M., 2022. Nipple-sparing

mastectomy with primary breast reconstruction: Breast cancer local recurrence according to molecular subtype. European Journal of Cancer,175, pp.S26-S27.

Goytia, G.I.C., Borges, R.H.M., Gonzalez, M.J. and Aponte, M.J.B., 2022. Journal of Cancer Therapy and Research.

Grimm, L.J., Avery, C.S., Hendrick, E. and Baker, J.A., 2022. Benefits and risks of mammography screening in women ages 40 to 49 years. *Journal of Primary Care & Community Health*, *13*, pp.21501327211058322.

Grootes, I., Keeman, R., Blows, F.M., Milne, R.L., Giles, G.G., Swerdlow, A.J., Fasching, P.A., Abubakar, M., Andrulis, I.L., Anton-Culver, H. and Beckmann, M.W., 2022. Incorporating progesterone receptor expression into the PREDICT breast prognostic model. European Journal of Cancer, 173, pp.178-193.

Hammond, M.E.H., Hayes, D.F., Dowsett, M., Allred, D.C., Hagerty, K.L., Badve, S., Fitzgibbons, P.L., Francis, G., Goldstein, N.S., Hayes, M. and Hicks, D.G., 2010. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Journal of Clinical Oncology*, *28*(16), pp.2784-2795.

Han, Y., Wang, J., Sun, Y., Yu, P., Yuan, P., Ma, F., Fan, Y., Luo, Y., Zhang, P., Li, Q. and Cai, R., 2021. Prognostic model and nomogram for estimating survival of small breast cancer: A SEER-based analysis. Clinical Breast Cancer, 21(5), pp.e497-e505.

Hanekom, S., 2018. The importance of ethical clearance in research. *Food For Thought, Research*.

Harper, J., Phillips, S., Munn, C., Vaughan, L., Biswakarma, R., Yasmin, E., Davies, M. and Talaulikar, V., 2023. P-568 Why menopause education is needed. Human Reproduction, 38(Supplement_1), pp.dead093-902.

Harrington,D.P and Fleming, T.R.(1982) A class of rank test procedures for censored survival data. Biometrika 69, pp 133-143

Hashmi, A.A., Naz, S., Hashmi, S.K., Hussain, Z.F., Irfan, M., Bakar, S.M.A., Faridi, N., Khan, A. and Edhi, M.M., 2018. Cytokeratin 5/6 and cytokeratin 8/18 expression in triple negative breast cancers: clinicopathologic significance in South-Asian population. BMC research notes, 11, pp.1-8.

Hashmi, A.A., Naz, S., Hashmi, S.K., Hussain, Z.F., Irfan, M., Bakar, S.M.A., Faridi, N., Khan, A. and Edhi, M.M., 2018. Cytokeratin 5/6 and cytokeratin 8/18 expression in triple negative breast cancers: clinicopathologic significance in South-Asian population. BMC research notes, 11, pp.1-8.

Hassen, A.M., Hussien, F.M., Asfaw, Z.A. and Assen, H.E., 2021. Factors associated with delay in breast cancer presentation at the only oncology center in North East Ethiopia: a cross-sectional study. *Journal of Multidisciplinary Healthcare*, pp.681-694.

Hausman, D.M., 2019. What is cancer?. Perspectives in biology and medicine, 62(4), pp.778-784.

He, J., Fu, F., Wang, W., Xi, G., Guo, W., Zheng, L., Ren, W., Qiu, L., Huang, X., Wang, C. and Li, L., 2021. Prognostic value of tumour-infiltrating lymphocytes based on the evaluation of frequency in patients with oestrogen receptor–positive breast cancer. European Journal of Cancer, 154, pp.217-226.

He, W., Zeng, E., Sjölander, A., Hübbert, L., Hedayati, E. and Czene, K., 2023. Concomitant Discontinuation of Cardiovascular Therapy and Adjuvant Hormone Therapy Among Patients With Breast Cancer. JAMA Network Open, 6(7), pp.e2323752-e2323752.

Honma, N., Ogata, H., Yamada, A., Matsuda, Y., Kontani, K., Miyashita, M., Arai, T., Sasaki, E., Shibuya, K., Mikami, T. and Sawaki, M., 2021. Clinicopathological characteristics and prognostic marker of triple-negative breast cancer in older women. Human Pathology, 111, pp.10-20.

Hortobagyi, G.N., de la Garza Salazar, J., Pritchard, K., Amadori, D., Haidinger, R., Hudis, C.A., Khaled, H., Liu, M.C., Martin, M., Namer, M. and O'Shaughnessy, J.A., 2005. The global breast cancer burden: variations in epidemiology and survival. Clinical breast cancer, 6(5), pp.391-401.

Hwang, K.T., Kim, Y.A., Kim, J., Oh, H.J., Park, J.H., Choi, I.S., Park, J.H., Oh, S., Chu, A., Lee, J.Y. and Hwang, K.R., 2021. Prognostic influences of BCL1 and BCL2 expression on disease-free survival in breast cancer. Scientific Reports, 11(1), pp.11942.

Jenkins, J.A., Marmor, S., Hui, J.Y.C., Beckwith, H., Blaes, A.H., Potter, D. and Tuttle, T.M., 2022. The 70-gene signature test as a prognostic and predictive biomarker in patients with invasive lobular breast cancer. Breast Cancer Research and Treatment, pp.1-7.

Ji, P., Gong, Y., Jin, M.L., Hu, X., Di, G.H. and Shao, Z.M., 2020. The burden and trends of breast cancer from 1990 to 2017 at the global, regional, and national levels: results from the global burden of disease study 2017. Frontiers in oncology, 10, pp.650.

Jia, T., Lv, Q., Cai, X., Ge, S., Sang, S., Zhang, B., Yu, C. and Deng, S., 2023. Radiomic signatures based on pretreatment 18F-FDG PET/CT, combined with clinicopathological characteristics, as early prognostic biomarkers among patients with invasive breast cancer. Frontiers in Oncology, 13, pp.1210125.

Johansson, A.L., Trewin, C.B., Hjerkind, K.V., Ellingjord-Dale, M., Johannesen, T.B. and Ursin, G., 2019. Breast cancer-specific survival by clinical subtype after 7 years follow-up of young and elderly women in a nationwide cohort. *International journal of cancer*, *144*(6), pp.1251-1261.

Joko-Fru, W.Y., Miranda-Filho, A., Soerjomataram, I., Egue, M., Akele-Akpo, M.T., N'da, G., Assefa, M., Buziba, N., Korir, A., Kamate, B. and Traore, C., 2020. Breast cancer survival in sub-Saharan Africa by age, stage at diagnosis and

human development index: A population-based registry study. *International journal of cancer*, *146*(5), pp.1208-1218

Joko-Fru, W.Y., Miranda-Filho, A., Soerjomataram, I., Egue, M., Akele-Akpo, M.T., N'da, G., Assefa, M., Buziba, N., Korir, A., Kamate, B. and Traore, C., 2020. Breast cancer survival in Sub-Saharan Africa by age, stage at diagnosis and human development index: A population-based registry study. International journal of cancer, 146(5), pp.1208-1218.

Jung, S.U., Sohn, G., Kim, J., Chung, I.Y., Lee, J.W., Kim, H.J., Ko, B.S., Son, B.H., Ahn, S.H., Yang, S.W. and Lee, S.B., 2019. Survival outcome of adjuvant endocrine therapy alone for patients with lymph node-positive, hormone-responsive, HER2-negative breast cancer. Asian journal of surgery, 42(10), pp.914-921.

Kalbfleisch, J. D. and Prentice, R. L. (2011). The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons.

Kalbfleisch,JD. And Prentice, RL. (1980). The statistical analysis of failure time data. John Wiley & Sons, New York, U.S.A.

Kaplan, E.L. and Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), pp.457-481.

Karutjaiva, M.S., 2021. *Breast cancer-associated risk factors and management among women attending Medical Imaging departments' in Namibia* (Doctoral dissertation, Namibia University of Science and Technology).

Karutjaiva, M.S., 2021. Breast cancer-associated risk factors and management among women attending Medical Imaging departments' in Namibia (Doctoral dissertation, Namibia University of Science and Technology).

Kast, K., John, E.M., Hopper, J.L., Andrieu, N., Noguès, C., Mouret-Fourme, E., Lasset, C., Fricker, J.P., Berthet, P., Mari, V. and Salle, L., 2023. Associations of height, body mass index, and weight gain with breast cancer risk in carriers of a pathogenic variant in BRCA1 or BRCA2: the BRCA1 and BRCA2 Cohort Consortium. Breast Cancer Research, 25(1), pp.1-13.

Kaur, S., Mayanglambam, P., Bajwan, D. and Thakur, N., 2022. Chemotherapy and its adverse effects-A systematic review. International Journal of Nursing Education and Research, 10(4), pp.399-402.

Kay, R., (2002). On the use of the accelated failure time model as an alternative to the proportional hazards models in the treatment of time to event data: a case study of influenza. Drug information Journal, 36, pp. 571-579.

Kazmi, S., Chatterjee, D., Raju, D., Hauser, R. and Kaufman, P.A., 2020. Overall survival analysis in patients with metastatic breast cancer and liver or lung metastases treated with eribulin, gemcitabine, or capecitabine. *Breast Cancer Research and Treatment*, *184*, pp.559-565.

Kensler, K.H., Poole, E.M., Heng, Y.J., Collins, L.C., Glass, B., Beck, A.H., Hazra, A., Rosner, B.A., Eliassen, A.H., Hankinson, S.E. and Winer, E.P., 2019. Androgen receptor expression and breast cancer survival: results from the nurses' health studies. JNCI: Journal of the National Cancer Institute, 111(7), pp.700-708.

Killian, M., Mahony, D.O., Murphy, K., Connor, D.O., Bird, B. and Murphy, C.G., 2023. Breast cancer outcomes in a private hospital appear better than national outcomes in a country with a mixed public/private healthcare model. Irish Journal of Medical Science (1971-), 192(2), pp.527-531.

Kim, H.S., Lee, J.U., Yoo, T.K., Chae, B.J., Son, D., Kim, Y.J. and Park, W.C., 2019. Omission of chemotherapy for the treatment of mucinous breast cancer: a

nationwide study from the Korean Breast Cancer Society. Journal of breast cancer, 22(4), pp.599-612.

Kizy, S., Altman, A.M., Marmor, S., Denbo, J.W., Jensen, E.H., Tuttle, T.M. and Hui, J.Y.C., 2019. 21-gene recurrence score testing in the older population with estrogen receptor-positive breast cancer. Journal of geriatric oncology, 10(2), pp.322-329.

Kohler, R.E., Gopal, S., Miller, A.R., Lee, C.N., Reeve, B.B., Weiner, B.J. and Wheeler, S.B., 2017. A framework for improving early detection of breast cancer in sub-Saharan Africa: A qualitative study of help-seeking behaviors among Malawian women. *Patient education and counseling*, *100*(1), pp.167-173.

Konduri, S., Singh, M., Bobustuc, G., Rovin, R. and Kassam, A., 2020. Epidemiology of male breast cancer. The Breast, 54, pp.8-14.

Kumar, A. and Jagannathan, N., 2018. Cytokeratin: A review on current concepts. International Journal of Orofacial Biology, 2(1), pp.6.

Lambert, M., Mendenhall, E., Kim, A.W., Cubasch, H., Joffe, M. and Norris, S.A., 2020. Health system experiences of breast cancer survivors in urban South Africa. *Women's Health*, *16*, pp.1745506520949419.

Lei, J., Wang, Y., Bi, Z., Xue, S., Ou, B. and Liu, K., 2020. Intraoperative radiotherapy (IORT) versus whole-breast external beam radiotherapy (EBRT) in early stage breast cancer: results from SEER database. Japanese journal of radiology, 38, pp.85-92.

Leone, J.P., Leone, B.A., Tayob, N., Hassett, M.J., Leone, J., Freedman, R.A., Tolaney, S.M., Winer, E.P., Vallejo, C.T. and Lin, N.U., 2021. Twenty-year risks of breast cancer-specific mortality for stage III breast cancer in the surveillance, epidemiology, and end results registry. Breast Cancer Research and Treatment, 187, pp.843-852.

Li, G.Q., Xie, S.J., Wu, S.G. and He, Z.Y., 2023. Impact of the 21-gene expression assay on treatment decisions and clinical outcomes in breast cancer with one to three positive lymph nodes. Frontiers in Endocrinology, 14, pp.1103949.

Li, M., Zhang, Y., Pei, L., Zhang, Z., Tan, G. and Huang, Y., 2022. Potential influence of anesthetic interventions on breast cancer early recurrence according to estrogen receptor expression: a sub-study of a randomized trial. Frontiers in Oncology, 12, pp.837959.

Li, S., Wu, Y., Ding, F., Yang, J., Li, J., Gao, X., Zhang, C. and Feng, J., 2020. Engineering macrophage-derived exosomes for targeted chemotherapy of triple-negative breast cancer. Nanoscale, 12(19), pp.10854-10862.

Li, X., Sanz, J., Foro, P., Martínez, A., Zhao, M., Reig, A., Liu, F., Huang, Y., Membrive, I., Algara, M. and Rodríguez, N., 2021. Long-term results of a randomized partial irradiation trial compared to whole breast irradiation in the early stage and low-risk breast cancer patients after conservative surgery. Clinical and Translational Oncology, 23, pp.2127-2132.

Li, Y., Chen, M., Pardini, B., Dragomir, M.P., Lucci, A. and Calin, G.A., 2019. The role of radiotherapy in metaplastic breast cancer: a propensity score-matched analysis of the SEER database. Journal of translational medicine, 17, pp.1-12.

Li, Y., Li, W., Gong, C., Zheng, Y., Ouyang, Q., Xie, N., Qu, Q., Ge, R. and Wang, B., 2021. A multicenter analysis of treatment patterns and clinical outcomes of subsequent therapies after progression on palbociclib in HR+/HER2− metastatic breast cancer. Therapeutic Advances in Medical Oncology, 13, pp.17588359211022890.

Lian, C.L., Cai, X.Y., Zhou, P., Wang, J., Chen, X.B. and Wu, S.G., 2020. Integration the biologic factors into the staging of breast cancer patients with ipsilateral supraclavicular lymph node metastasis. Journal of Cancer, 11(23), pp.6834.

Liang, Q., Ma, D., Gao, R.F. and Yu, K.D., 2020. Effect of Ki-67 expression levels and histological grade on breast cancer early relapse in patients with different immunohistochemical-based subtypes. Scientific reports, 10(1), pp.7648.

Liu, G., Xing, Z., Guo, C., Dai, Q., Cheng, H., Wang, X., Tang, Y. and Wang, Y., 2023. Identifying clinicopathological risk factors for regional lymph node metastasis in Chinese patients with T1 breast cancer: A population-based study. Frontiers in Oncology, 13, pp.1217869.

Liu, H., Zhang, X., Zhang, S., Wang, X. and Yu, S., 2021. Factors associated with bone metastasis in breast cancer: A systematic review and meta-analysis. Annals of Palliative Medicine, 10(4), pp.4435452-4434452.

Liu, Y., He, M., Zuo, W.J., Hao, S., Wang, Z.H. and Shao, Z.M., 2021. Tumor size still impacts prognosis in breast cancer with extensive nodal involvement. Frontiers in Oncology, 11, pp.585613.

Liu, Y., He, M., Zuo, W.J., Hao, S., Wang, Z.H. and Shao, Z.M., 2021. Tumor size still impacts prognosis in breast cancer with extensive nodal involvement. Frontiers in Oncology, 11, pp.585613.

Lu, T., Shih, W. J., & Johnson, R. E. (2005)**.** *Survival analysis with time-dependent covariates: Application and impact in health studies*. *Statistics in Medicine*, 24**(**22), pp.3393-3405.

Luo, C., Zhong, X., Wang, Z., Wang, Y., Wang, Y., He, P., Peng, Q. and Zheng, H., 2019. Prognostic nomogram for patients with non-metastatic HER2 positive breast cancer in a prospective cohort. The International journal of biological markers, 34(1), pp.41-46.

Lynden-Bell, D., 1971. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, *155*(1), pp.95-118.

Sharma M.K. & Abebe E. (2019). *Determinants of Survival Time of Women with Breast Cancer. Archives of Oncology and Cancer Therapy*, 2(2), pp.26-39.

Ma, S.J., Gill, J., Yendamuri, K., Chatterjee, U., Waldman, O., Dunne-Jaffe, C., Fekrmandi, F., Shekher, R., Iovoli, A., Yao, S. and Oladeru, O.T., 2023. Association of progesterone receptor status with 21-gene recurrence score and survival among patients with estrogen receptor-positive breast cancer. BMC cancer, 23(1), pp.330.

Makar, W.S., 2019. Clinicopathological Characteristics and Survival of Triple-Negative Breast Cancer Patients: A single Institution Study from Egypt. Research in Oncology, 15(1), pp.31-34.

Malmgren, J.A., Calip, G.S., Atwood, M.K., Mayer, M. and Kaplan, H.G., 2020. Metastatic breast cancer survival improvement restricted by regional disparity: surveillance, epidemiology, and end results and institutional analysis: 1990 to 2011. Cancer, 126(2), pp.390-399.

Mao, X., Omeogu, C., Karanth, S., Joshi, A., Meernik, C., Wilson, L., Clark, A., Deveaux, A., He, C., Johnson, T. and Barton, K., 2023. Association of reproductive risk factors and breast cancer molecular subtypes: a systematic review and meta-analysis. BMC cancer, 23(1), pp.644.

Martin, M., Hegg, R., Kim, S.B., Schenker, M., Grecea, D., Garcia-Saenz, J.A., Papazisis, K., Ouyang, Q., Lacko, A., Oksuzoglu, B. and Reeves, J., 2022. Treatment with adjuvant abemaciclib plus endocrine therapy in patients with high-risk early breast cancer who received neoadjuvant chemotherapy: a prespecified analysis of the monarchE randomized clinical trial. JAMA oncology, 8(8), pp.1190-1194.

Mavaddat, N., Pharoah, P.D., Michailidou, K., Tyrer, J., Brook, M.N., Bolla, M.K., Wang, Q., Dennis, J., Dunning, A.M., Shah, M. and Luben, R., 2015. Prediction of breast cancer risk based on profiling with common genetic variants. *Journal of the National Cancer Institute*, *107*(5), pp.djv036.

McGuire, W.L., Clark, G.M., Dressler, L.G. and Owens, M.A., 1986. Role of steroid hormone receptors as prognostic factors in primary breast cancer. *NCI Monogr*, *1*, pp.19-23.

Mishra, P., Davies, D.A. and Albensi, B.C., 2023. The Interaction Between NF-κB and Estrogen in Alzheimer's Disease. Molecular Neurobiology, 60(3), pp.1515-1526.

Moghadami Fard, Z. and Gohari, M.R., 2012. Survival analysis of patients with breast cancer using the Aalen's additive hazard model. Journal of North Khorasan University of Medical Sciences, 3(5), pp.171-179.

Moldoveanu, D., Pravongviengkham, V., Best, G., Martínez, C., Hijal, T., Meguerditchian, A.N., Lajoie, M., Dumitra, S., Watson, I. and Meterissian, S., 2020. Dynamic neutrophil-to-lymphocyte ratio: a novel prognosis measure for triple-negative breast cancer. Annals of Surgical Oncology, 27, pp.4028-4034.

Momenimovahed, Z. and Salehiniya, H., 2017. Incidence, mortality and risk factors of cervical cancer in the world. *Biomedical Research and Therapy*, *4*(12), pp.1795-1811.

Momenimovahed, Z. and Salehiniya, H., 2019. Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, pp.151-164.

Montazeri, A., Vahdaninia, M., Harirchi, I., Harirchi, A.M., Sajadian, A., Khaleghi, F., Ebrahimi, M., Haghighat, S. and Jarvandi, S., 2008. Breast cancer in Iran: need for greater women awareness of warning signs and effective screening methods. *Asia Pacific family medicine*, *7*(1), pp.1-7.

Montemurro, F., Nuzzolese, I. and Ponzone, R., 2020. Neoadjuvant or adjuvant chemotherapy in early breast cancer?. *Expert Opinion on Pharmacotherapy*, *21*(9), pp.1071-1082.

Moos, P.J. and Fitzpatrick, F.A., 1998. Taxanes propagate apoptosis via two cell populations with distinctive cytological and molecular traits. *Cell Growth and*

Differentiation-Publication American Association for Cancer Research, *9*(8), pp.687.

Morante, Z., Ruiz, R., Araujo, J.M., Pinto, J.A., de la Cruz-Ku, G., Urrunaga-Pastor, D., Namuche, F., Flores, C., Mantilla, R., Luján, M.G. and Fuentes, H., 2021. Impact of the delayed initiation of adjuvant chemotherapy in the outcome of triple negative breast cancer. Clinical Breast Cancer, 21(3), pp.239-246.

Nabegu, A.A., Muhammad, M.U. and Abdullahi, N., 2023. SURVIVAL ANALYSIS OF BREAST CANCER PATIENTS IN KANO STATE. FUDMA JOURNAL OF SCIENCES, 7(2), pp.199-205.

Nabholtz, J.M. and Gligorov, J., 2005. The role of taxanes in the treatment of breast cancer. *Expert opinion on pharmacotherapy*, *6*(7), pp.1073-1094.

Narod, S.A., Metcalfe, K., Lynch, H.T., Ghadirian, P., Robidoux, A., Tung, N., Gaughan, E., Kim-Sing, C., Olopade, O.I., Foulkes, W.D. and Robson, M., 2013. Should all BRCA1 mutation carriers with stage I breast cancer receive chemotherapy?. Breast cancer research and treatment, 138, pp.273-279.

Nelson, D.R., Brown, J., Morikawa, A. and Method, M., 2022. Breast cancer-specific mortality in early breast cancer as defined by high-risk clinical and pathologic characteristics. Plos one, 17(2), pp.e0264637.

Nelson, W., 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), pp.945-966.

Niţă, I., Niţipir, C., Toma, Ş.A., Limbău, A.M., Pirvu, E., Bădărău, I.A., Suciu, I., Suciu, G. and Manolescu, L.S.C., 2022. Level of education, background and clinical stage as prognostic factors according to RMST function in patients with early and locally advanced breast cancer: A single institution experience from Romania. Medicine and Pharmacy Reports, 95(1), pp.31.

Nunnery, S.E., Mayer, I.A. and Balko, J.M., 2021. Triple-negative breast cancer: breast tumors with an identity crisis. The Cancer Journal, 27(1), pp.2-7.

Oakes, D., 1977. The asymptotic information in censored survival data. *Biometrika*, *64*(3), pp.441-448.

Olarewaju, S.O., Oyekunle, E.O. and Bamiro, A.O., 2019. Effect of sociodemographic variables on patient and diagnostic delay of breast cancer at the Foremost health care institution in Nigeria. *Journal of global oncology*, *5*, pp.1-8.

Orozco, J.I., Keller, J.K., Chang, S.C., Fancher, C.E. and Grumley, J.G., 2022. Impact of locoregional treatment on survival in young patients with early-stage breast cancer undergoing upfront surgery. Annals of Surgical Oncology, 29(10), pp.6299-6310.

Oudanonh, T., Nabi, H., Ennour-Idrissi, K., Lemieux, J. and Diorio, C., 2020. Progesterone receptor status modifies the association between body mass index and prognosis in women diagnosed with estrogen receptor positive breast cancer. International Journal of Cancer, 146(10), pp.2736-2745.

Pace, L.E., Mpunga, T., Hategekimana, V., Dusengimana, J.M.V., Habineza, H., Bigirimana, J.B., Mutumbira, C., Mpanumusingo, E., Ngiruwera, J.P., Tapela, N. and Amoroso, C., 2015. Delays in breast cancer presentation and diagnosis at two rural cancer referral centers in Rwanda. *The oncologist*, *20*(7), pp.780-788.

Park, J.H., Jonas, S.F., Bataillon, G., Criscitiello, C., Salgado, R., Loi, S., Viale, G., Lee, H.J., Dieci, M.V., Kim, S.B. and Vincent-Salomon, A., 2019. Prognostic value of tumor-infiltrating lymphocytes in patients with early-stage triple-negative breast cancers (TNBC) who did not receive adjuvant chemotherapy. Annals of oncology, 30(12), pp.1941-1949.

Parkin, D.M. and Fernández, L.M., 2006. Use of statistics to assess the global burden of breast cancer. The breast journal, 12, pp.S70-S80.

Pascoal, C., Carrascal, M.A., Barreira, D.F., Lourenço, R.A., Granjo, P., Grosso, A.R., Borralho, P., Braga, S. and Videira, P.A., 2023. Sialyl LewisX/A and Cytokeratin Crosstalk in Triple Negative Breast Cancer. Cancers, 15(3), pp.731.

Patel, K., Kay, R.,and Rowel, L.(2006). Comparing proportional hazards and accelerated failure time models: an Application in influenza. Pharmaceutical Statististics, 5, pp.213-224.

Penel, N., Adenis, A. and Bocci, G., 2012. Cyclophosphamide-based metronomic chemotherapy: after 10 years of experience, where do we stand and where are we going?. *Critical reviews in oncology/hematology*, *82*(1), pp.40-50.

Peto, R and Peto, J (1972). Asymptotically efficient rank invariant test procedures (with discussion). J.R. Stat. Soc.Ser.A135, pp185-206

Peto, R., Davies, C., Godwin, J., Gray, R., Pan, H.C., Clarke, M., Cutter, D., Darby, S., McGale, P., Taylor, C. and Wang, Y.C., 2012. Early Breast Cancer Trialists' Collaborative G. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*, *379*(9814), pp.432-44.

Proietti, C.J., Cenciarini, M.E. and Elizalde, P.V., 2018. Revisiting progesterone receptor (PR) actions in breast cancer: Insights into PR repressive functions. Steroids, 133, pp.75-81.

Pruitt, L.C., Odedina, S., Anetor, I., Mumuni, T., Oduntan, H., Ademola, A., Morhason-Bello, I.O., Ogundiran, T.O., Obajimi, M., Ojengbede, O.A. and Olopade, O.I., 2020. Breast cancer knowledge assessment of health workers in Ibadan, Southwest Nigeria. *JCO global oncology*, *6*, pp.387-394.

Saadatmand, S., Bretveld, R., Siesling, S. and Tilanus-Linthorst, M.M., 2016. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173,797 patients. *Nederlands tijdschrift voor geneeskunde*, *160*, pp.A9800-A9800.

Saini, A., Kumar, M., Bhatt, S., Saini, V. and Malik, A., 2020. Cancer causes and treatments. *International Journal of Pharmaceutical Sciences and Research*, *11*(7), pp.3121-3134.

Saini, A., Kumar, M., Bhatt, S., Saini, V. and Malik, A., 2020. Cancer causes and treatments. International Journal of Pharmaceutical Sciences and Research, 11(7), pp.3121-3134.

Sakafu, L.L., Philipo, G.S., Malichewe, C.V., Fundikira, L.S., Lwakatare, F.A., Van Loon, K., Mushi, B.P., DeBoer, R.J., Bialous, S.A. and Lee, A.Y., 2022. Delayed diagnostic evaluation of symptomatic breast cancer in sub-Saharan Africa: A qualitative study of Tanzanian women. *Plos one*, *17*(10), pp.e0275639.

Saridakis, A., Berger, E.R., Harigopal, M., Park, T., Horowitz, N., Le Blanc, J., Zanieski, G., Chagpar, A., Greenup, R., Golshan, M. and Lannin, D.R., 2021. Apocrine breast cancer: unique features of a predominantly triple-negative breast cancer. Annals of Surgical Oncology, 28(10), pp.5610-5616.

Saunders, J., 2016. Confidentiality. *Medicine*, *44*(10), pp.596-597.

Scheel, J.R., Anderson, S., Foerster, M., Galukande, M. and McCormack, V., 2018. Factors contributing to late-stage breast cancer presentation in sub-Saharan Africa. Current Breast Cancer Reports, 10, pp.142-147.

Schoenfeld, D.A. (1982). Partial residuals for the proportional hazards regression model. Biometrika 69, pp239-241.

Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics, pp.461-464.

Seely, J.M. and Alhassan, T., 2018. Screening for breast cancer in 2018—what should we be doing today?. *Current Oncology*, *25*(s1), pp.115-124.

Senyefia, B.A., Joseph, A., Christiana Cynthia, N. and Ouerfelli, N., 2022. A comparison between accelerated failure time models in analyzing the survival of breast cancer patients. *Journal of Cancer and Tumor International*, *12*(1), pp.16-28.

Shen, C., Li, N., Zhou, S., Stahl, K., Dodge, D. and Zhao, H., 2023. Endocrine therapy initiation and overall survival outcomes with omission of radiation therapy in older Medicare patients with early-stage hormone-receptor-positive breast cancer. Cancer medicine, 12(6), pp.6935-6944.

Shen, K., Yu, H., Xie, B., Meng, Q., Dong, C., Shen, K. and Zhou, H.B., 2023. Anticancer or carcinogenic? The role of estrogen receptor β in breast cancer progression. Pharmacology & Therapeutics, pp.108350.

Simms, L., Barraclough, H. and Govindan, R., 2013. Biostatistics primer: what a clinician ought to know—prognostic and predictive factors. Journal of Thoracic Oncology, 8(6), pp.808-813.

Simon, J., Chaix, M., Billa, O., Kamga, A.M., Roignot, P., Ladoire, S., Coutant, C., Arveux, P., Quantin, C. and Dabakuyo-Yonli, T.S., 2020. Survival in patients with HR+/HER2− metastatic breast cancer treated with initial endocrine therapy versus initial chemotherapy. A French population-based study. British Journal of Cancer, 123(7), pp.1071-1077.

Sinha, S., Bhatia, R., Narasimamurthy, M., Rayne, S. and Grover, S., 2022. Epidemiology of Breast Cancer Presentation in Botswana, South Africa, and the United States. Journal of Surgical Research, *279*, pp.533-539.

Slighoua, M., Amrati, F.E.Z., Chebaibi, M., Mahdi, I., Al Kamaly, O., El Ouahdani, K., Drioiche, A., Saleh, A. and Bousta, D., 2023. Quercetin and Ferulic Acid Elicit Estrogenic Activities In Vivo and In Silico. Molecules, 28(13), pp.5112.

Soerjomataram, I., Louwman, M.W., Ribot, J.G., Roukema, J.A. and Coebergh, J.W.W., 2008. An overview of prognostic factors for long-term survivors of breast cancer. *Breast cancer research and treatment*, *107*, pp.309-330.

Soerjomataram, I., Louwman, M.W., Ribot, J.G., Roukema, J.A. and Coebergh, J.W.W., 2008. An overview of prognostic factors for long-term survivors of breast cancer. *Breast cancer research and treatment*, *107*, pp.309-330.

Song, S.E., Seo, B.K., Cho, K.R., Woo, O.H., Ganeshan, B., Kim, E.S. and Cha, J., 2021. Prediction of inflammatory breast cancer survival outcomes using computed tomography-based texture analysis. Frontiers in Bioengineering and Biotechnology, 9, pp.695305.

Sparano, J.A., Wang, M., Martino, S., Jones, V., Perez, E.A., Saphner, T., Wolff, A.C., Sledge Jr, G.W., Wood, W.C. and Davidson, N.E., 2008. Weekly paclitaxel in the adjuvant treatment of breast cancer. *New England Journal of Medicine*, *358*(16), pp.1663-1671.

Stabellini, N., Cullen, J., Cao, L., Shanahan, J., Hamerschlak, N., Waite, K., Barnholtz-Sloan, J.S. and Montero, A.J., 2023. Racial disparities in breast cancer treatment patterns and treatment related adverse events. Scientific Reports, 13(1), pp.1233.

Stahl, K., Wong, W., Dodge, D., Brooks, A., McLaughlin, C., Olecki, E., Lewcun, J., Newport, K., Vasekar, M. and Shen, C., 2021. Benefits of surgical treatment of stage IV breast cancer for patients with known hormone receptor and HER2 status. Annals of Surgical Oncology, 28, pp.2646-2658.

Ströbele, L., Kantelhardt, E.J., Traoré Millogo, T.F.D., Sarigda, M., Wacker, J. and Grosse Frie, K., 2018. Prevalence of breast-related symptoms, health care seeking behaviour and diagnostic needs among women in Burkina Faso. *BMC public health*, *18*, pp.1-7

Sun, J., Kong, L., Mu, K., Jiang, X., Luo, R., Wu, Y. and Ren, C., 2023. Survival and prognostic factors in patients with de novo metastatic breast cancer according to estrogen receptor status: A retrospective study.

Sun, T., Wang, T., Li, X., Wang, H. and Mao, Y., 2023. Tumor-infiltrating lymphocytes provides recent survival information for early-stage HER2-low-positive breast cancer: a large cohort retrospective study. Frontiers in Oncology, 13, pp.1148228.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), pp.209-249.

Tamirisa, N., Lin, H., Shen, Y., Shaitelman, S.F., Karuturi, M.S., Giordano, S.H., Babiera, G. and Bedrosian, I., 2020. Association of chemotherapy with survival in elderly patients with multiple comorbidities and estrogen receptor–positive, node-positive breast cancer. JAMA oncology, 6(10), pp.1548-1554.

Tang, C., Zhang, B., Yang, Y., Lin, Z. and Liu, Y., 2023. Overexpression of ferritin light chain as a poor prognostic factor for breast cancer. Molecular Biology Reports, pp.1-13.

Tarone, R.E. and Ware, J.H. (1977). On distribution-free test for equality for survival distributions. Biometrika64, pp156-160

Tashakkori, A. and Teddlie, C., 2010. Putting the human back in ‘‘human research methodology’’: The researcher in mixed methods research. *Journal of mixed methods research*, *4*(4), pp.271-277.

Taskindoust, M., Thomas, S.M., Sammons, S.L., Fayanju, O.M., DiLalla, G., Hwang, E.S. and Plichta, J.K., 2021. Survival outcomes among patients with metastatic breast cancer: review of 47,000 patients. *Annals of surgical oncology*, *28*(12), pp.7441-7449.

Tesfay, B., Getinet, T. and 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), pp.1908648.

Tuwei, G. and Degu, A., 2021. Survival outcomes among human epidermal growth factor receptor 2-(HER2-) positive breast cancer patients at Kenyatta National Hospital. *International Journal of Breast Cancer*, *2021*(1), pp.3115727.

Völkel, C., De Wispelaere, N., Weidemann, S., Gorbokon, N., Lennartz, M., Luebke, A.M., Hube-Magg, C., Kluth, M., Fraune, C., Möller, K. and Bernreuther, C., 2022. Cytokeratin 5 and cytokeratin 6 expressions are unconnected in normal and cancerous tissues and have separate diagnostic implications. Virchows Archiv, pp.1-15.

Waks, A.G. and Winer, E.P., 2019. Breast cancer treatment: a review. *Jama*, *321*(3), pp.288-300.

Walsh, S.M., Zabor, E.C., Stempel, M., Morrow, M. and Gemignani, M.L., 2019. Does race predict survival for women with invasive breast cancer?. *Cancer*, *125*(18), pp.3139-3146.

Wang, H. and Mao, X., 2020. Evaluation of the efficacy of neoadjuvant chemotherapy for breast cancer. *Drug design, development and therapy*, pp.2423-2433.

Wang, Z., Liu, L., Li, Y., Song, Z.A., Jing, Y., Fan, Z. and Zhang, S., 2021. Analysis of CK5/6 and EGFR and its effect on prognosis of triple negative breast cancer. Frontiers in Oncology, 10, pp.575317.

Wei, L.J., 1992. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine, 11(14-15), pp.1871-1879.

Wilkinson, A.N., 2023. Demystifying breast cancer. Canadian Family Physician, 69(7), pp.473.

Wolff, A.C., Hammond, M.E.H., Schwartz, J.N., Hagerty, K.L., Allred, D.C., Cote, R.J., Dowsett, M., Fitzgibbons, P.L., Hanna, W.M., Langer, A. and McShane, L.M., 2006. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Journal of clinical oncology*, *25*(1), pp.118-145.

Wu, M., Zhao, T., Zhang, Q., Zhang, T., Wang, L. and Sun, G., 2023. Prognostic analysis of breast cancer in Xinjiang based on Cox proportional hazards model and two− step cluster method. Frontiers in Oncology, 12, pp.1044945.

Xu, J., Guo, X., Jing, M. and Sun, T., 2018. Prediction of tumor mutation burden in breast cancer based on the expression of ER, PR, HER-2, and Ki-67. OncoTargets and therapy, pp.2269-2275.

Yang, S.X., Hewitt, S.M. and Yu, J., 2022. Locoregional tumor burden and risk of mortality in metastatic breast cancer. NPJ Precision Oncology, 6(1), pp.22.

Yang, Z.Y., Chen, W.L., Wu, W.T., Lai, C.H., Ho, C.L. and Wang, C.C., 2022. Return to work and mortality in breast cancer survivors: A 11-year longitudinal study. International Journal of Environmental Research and Public Health, 19(21), pp.14418.

Yao, Y., Chu, Y., Xu, B., Hu, Q. and Song, Q., 2019. Radiotherapy after surgery has significant survival benefits for patients with triple-negative breast cancer. Cancer medicine, 8(2), pp.554-563.

Yi, X., Hu, S., Ma, M., Huang, D. and Zhang, Y., 2024. Effect of HER2-low expression on neoadjuvant efficacy in operable breast cancer. Clinical and Translational Oncology, 26(4), pp.880-890.

Yoon, K.H., Park, Y., Kang, E., Kim, E.K., Kim, J.H., Kim, S.H., Suh, K.J., Kim, S.M., Jang, M., La Yun, B. and Park, S.Y., 2022. Effect of estrogen receptor expression level and hormonal therapy on prognosis of early breast

cancer. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, *54*(4), pp.1081-1090.

Zanoni, D.K., Patel, S.G. and Shah, J.P., 2019. Changes in the 8th edition of the American Joint Committee on Cancer (AJCC) staging of head and neck cancer: rationale and implications. Current oncology reports, 21, pp.1-7.

Zhang, B., Zhang, Z., Gao, B., Zhang, F., Tian, L., Zeng, H. and Wang, S., 2023. Raman microspectroscopy based TNM staging and grading of breast cancer. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 285, pp.121937.

Zhang, L., Hsieh, M.C., Petkov, V., Yu, Q., Chiu, Y.W. and Wu, X.C., 2020. Trend and survival benefit of Oncotype DX use among female hormone receptor-positive breast cancer patients in 17 SEER registries, 2004–2015. Breast cancer research and treatment, 180, pp.491-501.

Appendix

R CODES

R STUDIO SYNTAXES/CODES FOR DESCRIPTIVE STATISTICS

```
install.packages("flexsurv")
install.packages("gtsummary")
install.packages("SurvRegCensCov")
install.packages("tidyverse")
install.packages('broom')
install.packages('dplyr')
install.packages('ggplot2')
install.packages('risk.table.height')
install.packages('survival')
install.packages('survminer')
install.packages('tidyr')
library(broom)
library(dplyr)
library(flexsurv)
library(ggplot2)
library(gtsummary)
library(survival)
library(survminer)
library(SurvRegCensCov)
library(tidyr)
library(tidyverse)
```

```
options(tibble.print_min = Inf)
options(tibble.width = Inf)
print(KM1a, n=Inf)
```

FITTING COX-SNELL RESIDUAL PLOT FOR DIFFERENT PARAMETRIC AFT MODEL

Exponential

```
fitg <- flexsurvreg(formula = Surv(Survival, Status) ~ 1, data = breast1, dist = "exponential")
cs <- coxsnell_flexsurvreg(fitg)
## Model appears to fit well, with some small sample noise
surv <- survfit(Surv(cs$est, breast1$Status) ~ 1)
plot(surv, fun="cumhaz",xlab = "Exponential Reg Model Cum Hazard", ylab = "Kaplan-Meier hazard",conf.int = FALSE,
    main = "Exponential AFT Model")
abline(0, 1, col="red")
```

Weibull

```
fitg <- flexsurvreg(formula = Surv(Survival, Status) ~ 1, data = breast1, dist = "weibull")
cs <- coxsnell_flexsurvreg(fitg)
## Model appears to fit well, with some small sample noise
surv <- survfit(Surv(cs$est, breast1$Status) ~ 1)
plot(surv, fun="cumhaz",xlab = "Weibull Reg Model Cum Hazard", ylab = "Kaplan-Meier hazard",conf.int = FALSE,
    main = "Weibull AFT Model")
abline(0, 1, col="red")
```

Log-logistic

```
fitg <- flexsurvreg(formula = Surv(Survival, Status) ~ 1, data = breast1, dist = "llogis")
cs <- coxsnell_flexsurvreg(fitg)
## Model appears to fit well, with some small sample noise
```

```
surv <- survfit(Surv(cs$est, breast1$Status) ~ 1)
plot(surv, fun="cumhaz",xlab = "Log-logistic Reg Model Cum Hazard", ylab = "Kaplan-Meier hazard",conf.int = FALSE,
    main = "Log-logistic AFT Model")
abline(0, 1, col="red")
```

log-normal

```
fitg <- flexsurvreg(formula = Surv(Survival, Status) ~ 1, data = breast1, dist = "lnorm")
cs <- coxsnell_flexsurvreg(fitg)

## Model appears to fit well, with some small sample noise
surv <- survfit(Surv(cs$est, breast1$Status) ~ 1)
plot(surv, fun="cumhaz",xlab = " Log-normal Reg Model Cum Hazard", ylab = "Kaplan-Meier hazard",conf.int = FALSE,
    main = "Log-normal AFT Model")
abline(0, 1, col="red")
```

Gamma

```
fitg <- flexsurvreg(formula = Surv(Survival, Status) ~ 1, data = breast1, dist = "gamma")
cs <- coxsnell_flexsurvreg(fitg)

## Model appears to fit well, with some small sample noise
surv <- survfit(Surv(cs$est, breast1$Status) ~ 1)
plot(surv, fun="cumhaz",xlab = "Generalised gamma Reg Model Cum Hazard", ylab = "Kaplan-Meier hazard",conf.int = FALSE,
    main = "Gamma distribution")
abline(0, 1, col="red")
```

Cox PH

```
fit1 <- coxph(Surv(Survival,Status)~1,method='breslow',data=breast1)
coxsnellres <- breast1$Status-resid(fit1,type="martingale")
## Use NA method to estimate the cumulative hazard function for residuals
Cox_PH <- survfit(Surv(coxsnellres,breast1$Status)~1)
Kaplan_Meier_hazard <- cumsum(Cox_PH$n.event/fit4$n.risk)
## plot the results
plot(Cox_PH$time,Kaplan_Meier_hazard,xlab = " Cox PH Model Cum Hazard", ylab
= "Kaplan-Meier hazard",
    main = "Cox PH",type='s',col='black')
abline(0,1,col='red')
```

#Calculations of number of covariates and parameters and AIC for Cox PH

```
coxph_model <- coxph(Surv(Survival, Status) ~ 1, data = breast1)
coxph_model
# Extracting the log-likelihood
log_likelihood <- cox_model$loglik[2]
# Number of covariates
num_covariates <- length(cox_model$coefficients)
num_covariates
# Number of parameters (including the baseline hazard)
num_parameters <- num_covariates + 1  # Typically +1 for the intercept or baseline
hazard
num_parameters
# Calculating AIC
AIC_value <- -2 * log_likelihood + 2 * num_parameters
# Display results
cat("Log-Likelihood:", log_likelihood, "\n")
cat("Number of Covariates:", num_covariates, "\n")
cat("Number of Parameters:", num_parameters, "\n")
cat("AIC:", AIC_value, "\n")
```

or

```
cox_model <- coxph(Surv(Survival, Status) ~ 1, data = breast)
# Extract the log partial likelihood from the cox model object
loglik <- cox_model$loglik[2]
# Calculate the number of parameters in the model (including the intercept)
num_parameters <- length(coef(cox_model))
# Calculate AIC
aic <- -2 * loglik + 2 * num_parameters
# Print AIC
print(aic)

wei.mod.aft1 <- survreg(Surv(Survival, Status) ~ Menopausal_status_+ER_+PR_+HER2_+CK56_+Subtype_+Grade_+T_stage_+N_stage_+M_stage_+AJCC_stage_+Surgery_+Chemotherapy_+ET_,data = breast1, dist = 'weibull')
summary(wei.mod.aft1)
# Number of covariates
num_covariatesweibull <- length(wei.mod.aft1$coefficients)
num_covariatesweibull
# Number of parameters (including the baseline hazard)
num_parameters <- num_covariatesweibull + 1  # Typically +1 for the intercept or baseline hazard
num_parameters

exponential.aft <- survreg(Surv(Survival, Status) ~ Menopausal_status_+ER_+PR_+HER2_+CK56_+Subtype_+Grade_+T_stage_+N_stage_+M_stage_+AJCC_stage_+Surgery_+Chemotherapy_+ET_,data = breast1, dist = 'exponential')
```

```
summary(exponential.aft)
# Number of covariates
num_covariatesexponential <- length(exponential.aft $coefficients)
num_covariatesexponential
# Number of parameters (including the baseline hazard)
num_parameters <- num_covariatesexponential + 1  # Typically +1 for the intercept or baseline hazard
num_parameters

lognormal.aft <- survreg(Surv(Survival, Status) ~ Menopausal_status_+ER_+PR_+HER2_+CK56_+Subtype_+Grade_+T_stage_+N_stage_+M_stage_+AJCC_stage_+Surgery_+Chemotherapy_+ET_,data = breast1, dist = 'lognormal')
summary(lognormal.aft)
# Number of covariates
num_covariateslognormal <- length(lognormal.aft $coefficients)
num_covariateslognormal
# Number of parameters (including the baseline hazard)
num_parameters <- num_covariateslognormal + 1  # Typically +1 for the intercept or baseline hazard
num_parameters

loglogistic.aft <- survreg(Surv(Survival, Status) ~ Menopausal_status_+ER_+PR_+HER2_+CK56_+Subtype_+Grade_+T_stage_+N_stage_+M_stage_+AJCC_stage_+Surgery_+Chemotherapy_+ET_,data = breast1, dist = 'loglogistic')
summary(loglogistic.aft)
# Number of covariates
num_covariatesloglogistic <- length(loglogistic.aft $coefficients)
num_covariatesloglogistic
```

```
# Number of parameters (including the baseline hazard)
num_parameters <- num_covariatesloglogistic + 1  # Typically +1 for the intercept or baseline hazard
num_parameters
```

FITTING COX-PH MODEL FOR BREAST CANCER PATIENT'S DATASET IN PIETERSBURG HOSPITAL IN LIMPOPO PROVINCE

```
cox_model <- coxph(Surv(Survival, Status) ~Age_+ Menopausal_status_ + ER_ + PR_ + HER2_ + CK56_ + Subtype_ + Grade_ + T_stage_ + N_stage_ + M_stage_ + AJCC_stage_ + Surgery_ + Chemotherapy_ + ET_, data = breast1)
# Extract coefficients
coefficients <- summary(cox_model)$coefficients
# Calculate hazard ratios and confidence intervals
hazard_ratios <- exp(coefficients[, "coef"])
conf_int <- exp(confint(cox_model))
# Combine into a data frame
results <- data.frame(
  Covariate = rownames(coefficients),
  Estimate = coefficients[, "coef"],
  `Std. Error` = coefficients[, "se(coef)"],
  `Hazard Ratio` = hazard_ratios,
  `CI Lower` = conf_int[, 1],
  `CI Upper` = conf_int[, 2],
  `p-value` = coefficients[, "Pr(>|z|)"])
# Print the results
print(results)


# Fit the survival model
```

```
KM2aa <- survfit(Surv(Survival, Status) ~ Menopausal_status_, data = breast1)

# Plot with ggsurvplot
KM2aa_plot <- ggsurvplot(
  KM2aa,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.9),
  legend.title = "+   Censored\n Log-rank  P=0.071",
  legend.labs = c("Postmenopausal", "Premenopausal")
)

# Display the plot
KM2aa_plot

# Fit the survival model
KM2ER <- survfit(Surv(Survival, Status) ~ ER_, data = breast1)

# Plot with ggsurvplot
KM2ER_plot <- ggsurvplot(
  KM2ER,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
```

```
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.9),
  legend.title = "+   Censored\n Log-rank  P=0.36",
  legend.labs = c("ER Negative", "ER Positive")
)

# Display the plot
KM2ER_plot




# Fit the survival model
KM2PR <- survfit(Surv(Survival, Status) ~ PR_, data = breast1)

# Plot with ggsurvplot
KM2PR_plot <- ggsurvplot(
  KM2PR,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.9),
  legend.title = "+   Censored\n Log-rank  P=0.31",
  legend.labs = c("PR Negative", "PR Positive")
```

```
)

# Display the plot
KM2PR_plot

# Fit the survival model
KM2HER2_ <- survfit(Surv(Survival, Status) ~ HER2_, data = breast1)

# Plot with ggsurvplot
KM2HER2_plot <- ggsurvplot(
  KM2HER2_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.9),
  legend.title = "+   Censored\n Log-rank  P=0.62",
  legend.labs = c("HER2 Negative", "HER2 Positive")
)

# Display the plot
KM2HER2_plot




# Fit the survival model
```

```
KM2CK56_ <- survfit(Surv(Survival, Status) ~ CK56_, data = breast1)

# Plot with ggsurvplot
KM2CK56_plot <- ggsurvplot(
  KM2CK56_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.9),
  legend.title = "+   Censored\n Log-rank  P=0.39",
  legend.labs = c("CK5/6 Negative", "CK5/6 Positive")
)

# Display the plot
KM2CK56_plot

# Fit the survival model
KM2Subtype_ <- survfit(Surv(Survival, Status) ~ Subtype_, data = breast1)

# Plot with ggsurvplot
KM2Subtype_plot <- ggsurvplot(
  KM2Subtype_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
```

```
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.41",
  legend.labs = c("Luminal-A","Luminal-B")
)

# Display the plot
KM2Subtype_plot

# Fit the survival model
KM2Grade_ <- survfit(Surv(Survival, Status) ~ Grade_, data = breast1)

# Plot with ggsurvplot
KM2Grade_plot <- ggsurvplot(
  KM2Grade_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.14",
  legend.labs = c("Grade-1", "Grade-2","Grade-3")
)

# Display the plot
```

KM2Grade_plot

```
# Fit the survival model
KM2T_stage_ <- survfit(Surv(Survival, Status) ~ T_stage_, data = breast1)

# Plot with ggsurvplot
KM2T_stage_plot <- ggsurvplot(
  KM2T_stage_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.00036",
  legend.labs = c("T1", "T2","T3","T4")
)

# Display the plot
KM2T_stage_plot

# Fit the survival model
KM2N_stage_ <- survfit(Surv(Survival, Status) ~ N_stage_, data = breast1)

# Plot with ggsurvplot
KM2N_stage_plot <- ggsurvplot(
  KM2N_stage_,
  ylab = "Survival Probability",
```

```
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.012",
  legend.labs = c("N0", "N1","N2","N3")
)

# Display the plot
KM2N_stage_plot

# Fit the survival model
KM2N_stage_ <- survfit(Surv(Survival, Status) ~ N_stage_, data = breast1)

# Plot with ggsurvplot
KM2N_stage_plot <- ggsurvplot(
  KM2N_stage_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.012",
  legend.labs = c("N0", "N1","N2","N3","Nx")
```

```
)

# Display the plot
KM2N_stage_plot

# Fit the survival model
KM2M_stage_ <- survfit(Surv(Survival, Status) ~ M_stage_, data = breast1)

# Plot with ggsurvplot
KM2M_stage_plot <- ggsurvplot(
  KM2M_stage_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.0001",
  legend.labs = c("M0", "M1")
)

# Display the plot
KM2M_stage_plot

AJCC_stage_
# Fit the survival model
KM2AJCC_stage_ <- survfit(Surv(Survival, Status) ~ AJCC_stage_, data = breast1)
```

```
# Plot with ggsurvplot
KM2AJCC_stage_plot <- ggsurvplot(
  KM2AJCC_stage_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.0001",
  legend.labs = c("Stage-1", "Stage-2","Stage-3","Stage-4")
)

# Display the plot
KM2AJCC_stage_plot
```

```
Surgery_

# Fit the survival model
KM2Surgery_  <- survfit(Surv(Survival, Status) ~ Surgery_ , data = breast1)

# Plot with ggsurvplot
KM2Surgery_plot <- ggsurvplot(
  KM2Surgery_,
  ylab = "Survival Probability",
```

```
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.0001",
  legend.labs = c("Surgery=No", "Surgery=Yes")
)

# Display the plot
KM2Surgery_plot

Chemotherapy_
# Fit the survival model
KM2Chemotherapy_ <- survfit(Surv(Survival, Status) ~ Chemotherapy_, data = breast1)

# Plot with ggsurvplot
KM2Chemotherapy_plot <- ggsurvplot(
  KM2Chemotherapy_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
```

```
  legend.title = "+   Censored\n Log-rank  P=0.55",
  legend.labs = c("Chemotherapy=No", "Chemotherapy=Yes")
)
# Display the plot
KM2Chemotherapy_plot

ET_
KM2ET_ <- survfit(Surv(Survival, Status) ~ ET_, data = breast1)

# Plot with ggsurvplot
KM2ET_plot <- ggsurvplot(
  KM2ET_,
  ylab = "Survival Probability",
  xlab = "Time (Months)",
  data = breast1,
  pval = TRUE,
  surv.median.line = "hv",
  conf.int = FALSE,
  xlim = c(0, 120),
  legend = c(0.7, 0.8),
  legend.title = "+   Censored\n Log-rank  P=0.0018",
  legend.labs = c("Endocrine therapy=No", "Endocrine therapy=Yes")

# Display the plot
KM2ET_plot

#Median table for all variables in the breast cancer dataset

KM1aa<-survfit(Surv(Survival,Status)~1,data=breast1) # code to Calculate the med
ian of all BC patients
```

```
KM1aa          #Print the median table for all BC patients


# Median of Menopausal_status
KM2a<-survfit(Surv(Survival,Status)~ Menopausal status,data=breast1)
KM2          # Print the median table for Menopausal_status

#Median of ER
KM3a<-survfit(Surv(Survival,Status)~ER,data=breast1)
KM3a         # Print the median table for ER

#Median of PR
KM4a<-survfit(Surv(Survival,Status)~PR,data=breast1)
KM4a         # Print the median table for PR

#Median of HER2
KM5a<-survfit(Surv(Survival,Status)~HER2,data=breast1)
KM5a         # Print the median table for HER2

#Median of CK5/6
KM6a<-survfit(Surv(Survival,Status)~CK56,data=breast1)
KM6a         # Print the median table for CK5/6

#Median of Subtype
KM7a<-survfit(Surv(Survival,Status)~Subtype,data=breast1)
KM7a         # Print the median table for Subtype

#Median of Grade
KM8a<-survfit(Surv(Survival,Status)~Grade,data=breast1)
KM8a         # Print the median table for Grade
```

```
#Median of T stage
KM9a<-survfit(Surv(Survival,Status)~T_stage,data=breast1)
KM9a          # Print the median table for T stage

#Median of N stage
KM10a<-survfit(Surv(Survival,Status)~N_stage,data=breast1)
KM10a                # Print the median table for N stage

#Median of M stage
KM11a<-survfit(Surv(Survival,Status)~M_stage,data=breast1)
KM11a                # Print the median table for M stage

#Median of AJCC stage
KM12a<-survfit(Surv(Survival,Status)~AJCC_stage,data=breast1)
KM12a                # Print the median table for AJCC stage

#Median of Surgery
KM13a<-survfit(Surv(Survival,Status)~Surgery,data=breast1)
KM13a                # Print the median table for Surgery

#Median of Chemotherapy
KM14a<-survfit(Surv(Survival,Status)~ Chemotherapy,data=breast1)
KM14a                # Print the median table for Chemotherapy

#Median of ET
KM15a<-survfit(Surv(Survival,Status)~ ET,data=breast1)
KM15a                # Print the median table for ET
```

Predicted Probability Diagnostics

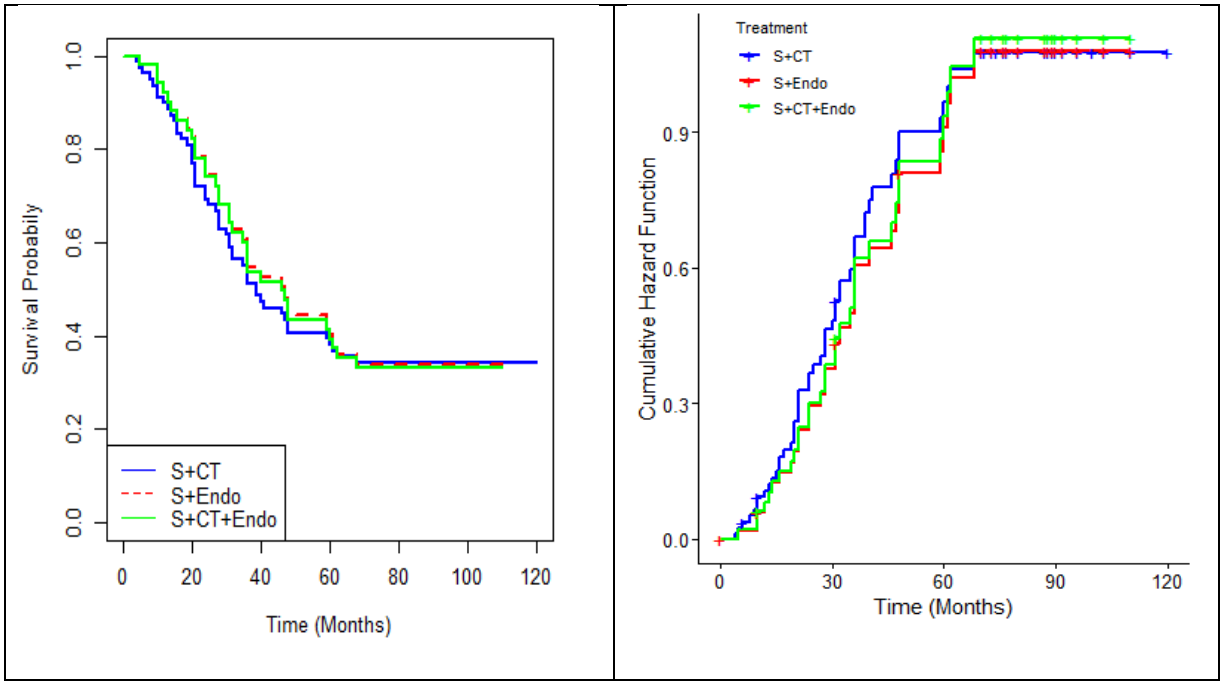


Figure 5. : Predicted Probability Diagnostics Plots