



UNIVERSITY OF NAIROBI

**DETERMINE RISK FACTORS OF BREAST CANCER IN
MOGADISHU**

BY

MAHAD GELLE SALAD

I56/33069/2019

**A Dissertation Submitted for Examination in Partial Fulfillment of the
Requirements for Award of the Degree of Master of Science in Social
Statistics of the University of Nairobi**

2021

Abstract

Cancer is a sickness that affects people of all socioeconomic backgrounds and cultures in the same way. Cancer will make treatment less likely to succeed for the remission of the patients and decrease survival chances if it is not diagnosed and treated promptly.

The goal of the research was to determine or look at risk factors of breast cancer in Mogadishu, Somalia regarding females only as well as evaluating the breast cancer prevalence. For the years February 2015 up to March 2020, secondary data was collected from the Osman hospital in Mogadishu, but we generalized the data from the whole country because the hospitals that can deal with breast cancer were located in Mogadishu, Somalia.

Descriptive statistics were calculated in the form of graphs and tables copied from R studio or excel platforms to search for risk factors of breast cancer. The time from diagnosis of breast cancer to death of the female patient is represented by a time to event variable. The survival rate of breast cancer patients was calculated using survival methodological approaches including the Kaplan-Meier method (Kaplan-Meier is a common method for dealing with this problem since it re-estimates the survival probability each time an occurrence occurs), log-rank test (The log-rank test is a widely used method for determining if two or more independent groups have the same chance of survival or not) and cox proportional hazard model (Cox proportional hazards regression is one of the most widely used regression methods for survival analysis. It is used to link multiple risk factors or exposures, all of which are considered at the same time, to survival time).

All risk factors from the data were found to be a strong predictor of breast cancer since they were all statistically significant in the study. The Hazard Ratio will increase as women patients get older, but the Hazard Ratio will decrease for women patients who received chemotherapy. The best way to fight breast cancer is using chemotherapy. The Hazard Ratio will increase for obese women patients, so fat women have a higher risk than others for breast cancer.


The study recommends that a greater focus on the breast cancer treatment passageway in Somalia and the creation of a Somalia National Cancer Registry in Mogadishu and other large towns in Somalia's sub-counties.

Keywords: Breast cancer, Kaplan Meier, Log rank and Cox Proportional Hazard

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics
©Mahad Gelle Salad, 2021
DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.



Signature

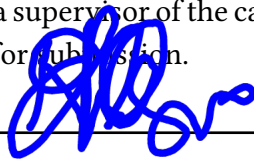
August 23, 2021

Date

MAHAD GELLE SALAD

Reg No. I56/33069/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.




Signature

August 23, 2021

Date

Dr. Nelson Onyango
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: onyango@uonbi.ac.ke



Signature

August 24, 2021

Date

Dr. Rachel Sarguta
School of Mathematics
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: rsarguta@uonbi.ac.ke

Dedication

To my Sisters (Nasteho Gelle Salad and Barlin Gelle Salad), my mother (Halima Mohamed Mohamud) and my cousins (Naima Abdi Jama, Ahsa Xassan Salad and Liban Axmed Abada). Thank you for your love and support.

Contents

Abstract	ii
Declaration and Approval	v
Dedication	viii
Figures and Tables	xi
Acknowledgments	xii
1 Introduction	1
1.1 Background.....	1
1.2 Statement of Problem.....	3
1.3 Objectives.....	4
1.4 Justification/Significance of Study.....	4
1.5 Structure of the thesis.....	5
2 Literature Review	6
2.1 Breast Cancer.....	6
2.2 Survival analysis.....	7
2.3 Overview of Literature Review.....	8
3 Research Methods	9
3.1 Research Design.....	9
3.2 Population of the Study.....	9
3.3 Sample.....	9
3.4 Procedure.....	9
3.5 Data analysis.....	10
3.6 Survival analysis.....	10
3.6.1 Censoring.....	11
3.6.2 Survival Function: $S(t)$	11
3.7 Kaplan Meier Estimator.....	13
3.7.1 Kaplan-Meier Estimator for the Survival Function.....	14
3.8 Log-Rank Test.....	15
3.9 Cox Proportional Hazard.....	16
3.9.1 Hazard Function, $h(t)$	16
3.9.2 Cox Proportional Hazard Model.....	17
4 Data Analysis And Results	20
4.1 Introduction.....	20
4.1.1 Check for violation of proportional hazard.....	20
4.2 Descriptive statistics.....	23
4.3 Survival Analysis.....	29
4.3.1 Kaplan Meier (KM).....	30
4.3.2 Cox Proportional Hazard model.....	32
5 Conclusion And Recommendations	37
5.1 Summary.....	37
5.2 Conclusion.....	37
5.3 Recommendations.....	38
5.4 Further Research.....	38

Bibliography..... 39

Figures and Tables

Figures

Figure 1. A visualization for the meaning of Kaplan-Meier table	15
Figure 2. Cox Proportional Hazard assumptions for breast cancer data	20
Figure 3. The use of graphs to check for time independence shows us whether the variable is batten tome or not	21
Figure 4. Variables' outlier for breast cancer data from the cox proportional model.....	21
Figure 5. Using breast cancer data from the Cox proportional model to test for age linearity, which is a continuous variable	22
Figure 6. Using breast cancer data from the cox proportional model to test linearity for a continuous variable, age at first birth.....	22
Figure 7. summary for the Category variables in the breast cancer data	24
Figure 8. Relationship between chemotherapy with tumor grade and breast	28
Figure 9. Relationship between chemotherapy and obesity and also between obesity and tumor grade .	29
Figure 10. Women breast cancer data structure from Mogadishu, Somalia	30
Figure 11. Kaplan Meier table.....	31
Figure 12. the kaplan Meier curve for chemotherapy and Obesity groups.....	31
Figure 13. summary for log rank test for Chemotherapy and Obesity groups	32
Figure 14. Plots of kaplan Meier for Tumor Grade and Breast	32
Figure 15. log rank test summary for tumor grade and breast groups.....	33
Figure 16. model selection from breast cancer data from the cox proportional model.....	33
Figure 17. Good fit model from breast cancer data from the cox proportional model.....	34
Figure 18. the results for the cox proportional model from breast cancer data	34

Tables

Table 1. summary of descriptive statistics for the breast cancer data with contentious variables	23
Table 2. In the breast cancer data, the average time of women who are obese and have breast groups on the left or right	25
Table 3. The average tumor size in breast cancer data for women with left or right breast cancer, whether they were receiving chemotherapy or not.....	26
Table 4. The average of age in breast cancer data for women with left or right breast cancer, whether they were receiving chemotherapy or not, plus obesity	27
Table 5. A hazard ratio for the Cox Proportional Hazard Model result is shown in the table	35

Acknowledgments

I am thankful to God Almighty for providing me with the grace, good health, ability, and strength to complete my work. I would want to offer my sincere appreciation, gratitude and admiration to my supervisors DR Nelson Onyango and DR Rachel Sarguta for their wise advice, invaluable assistance, passion, patience, and unwavering support. Their generosity and emotional support have served as a constant source of motivation. In fact, there aren't enough words to express my thankfulness to them.

I would want to express my gratitude to Dr. Nelson again and my committee for their insightful comments on how to improve the presentation of this thesis. A special thanks also to Professor J.A.M Ottieno for giving me his Survival Book and for assisting me when I needed it, I will remember his honesty and kindness.

Thank you to my classmates for supporting me throughout the development of this project. Thank you to all the lecturers who taught me, and gave me the knowledge that I was able to complete my thesis project and my master's course.

Finally, I wish to express my deep gratitude to my parents, cousins, my friends, and, above all, for their kind support, generous love and moral encouragement.

Mahad Gelle Salad

Nairobi, 2021.

1 Introduction

1.1 Background

Given that the world bank divides income growth in countries into two groups, developed countries with high income and developing countries with low or middle income, such categorization is critical for determining how countries' limited funds should be distributed to address the world's most pressing health problems, around 1.67 million new cases were reported in 2012, breast cancer is the highly common cancer in women around the world, breast cancer is also the first killer cancer in women in the world, accounting for roughly 500000 deaths per year, Breast cancer is now the most prevalent disease in countries of all income levels, and it is rapidly growing, with approximately 2.4 million cases reported in 2015 compared to 1.7 million in 2005 [Unger-Saldaña, et al., 2014]. Also, there are many problems in Africa and most developing countries, including inadequate health systems and unfinished vital registers, breast cancer mortality is high due to a lack of population understanding of breast cancer, poor health research activity, and low rates of female education and how to treat the disease, cancer was the leading cause of sickness and loss of life among women worldwide in 2015, with nearly 17.5 million cancer cases and 9 million deaths [Kantelhardt, et al., 2018].

According to WHO statistics from 2015, cancer was the first or second cause of death in 91 of 172 countries before the age of 70, and the third or fourth cause of death in 22 countries, the current global demographic shifts would result in an increased cancer burden, specifically in low and middle-income countries, over the next few decades, in 2012, the global cancer burden was reported to have resulted in 14.1 million new cases and 8.2 million deaths, rising to 18.1 million new cases and 9.6 million death in 2018, one for every five men and one for every six women will develop cancer over their lifetime, with one for every eight men and one for every eleven women dying from it, in 2018, Africa is projected to see 1,055,172 new cancer cases (5.8% of global total) and 693,487 cancer deaths (7.3%), the world has been divided into 20 regions to estimate cancer incidence and mortality rates, and the East African region includes Somali [Tahtabasi, et al., 2020].

Somalia is located in the horn of Africa, also known as eastern Africa, according to the Eastern Africa breast cancer analysis, around 133 900 new breast cancer cases in women were recorded in 2012 from Africa, accounting for 27.6% of all cancer cases, between 1991 and 2010, the age-standardized incidence rates per 100 000 women with breast cancer increased by 4.9 and 3.7 percent per year, respectively, from 20.9 to 46.8 and 18.0 to 31.2 new cases, in East African countries. Breast disease is the most widespread cancer in women in Northern Africa, East Africa, and several Sub-Saharan African countries, according to population-based cancer registries, Sudan has successfully conducted a report on timely identification by qualified participants, supportive care, involving pain man-

agement, was recently emphasized during and after adjunctive therapy for metastatic breast cancer, health initiatives should remedy this, and resources ought to be available starting at the most fundamental level [Kantelhardt, et al., 2015].

The following are the most common cancers affecting people in East African countries, in terms of prevalence and incidence among women or both sexes, taking into account the incidence of Cervix uteri cancer 54560 new cases (26.7%), breast cancer 45709 new cases (22.4%), colorectum cancer 9418 new cases (4.6%), oesophage 7623 new cases (3.7%), and ovary cancer 7298 new cases (3.7%). [<https://gco.iarc.fr/today>, Globocan, 2020]

The prevalence of breast cancer in east African countries for both sexes is Cervix uteri cancer 98660 cases (16.7%), breast cancer 88696 cases (15%), which means it is the second cancer after Cervix uteri cancer in terms of total cancers 590533, prostate 37669 cases (6.4%), Colorectum cancer 31317 cases (5.3%), and Kaposi sarcoma cancer 31011 cases (5.3%) considering cohort study in 5 years in 2020. [<https://gco.iarc.fr/today>, Globocan, 2020]

Comparing mortality between Somalia and East Africa for cancer cases the first killer in East Africa is Cervix uteri cancer for both sexes, while breast cancer is the first deadly cancer in Somalia, although Cervix uteri cancer is the second [<https://gco.iarc.fr/today>, Globocan, 2020].

No cancer register office in Somalia can keep data set and information about cancer, so, the first research to examine the cancer incidence in Somalia was done in 2017 for the city of Mogadishu and its environs, the findings or the available data were insufficient to illustrate the actual truth for everyone in the population because of the small number of patients, but the cancer incidence research in the Somalia population was performed with immigrants living in the United States of America (USA), and the majority of them only looked at women and a particular form of cancer, such as cervical or breast cancer, the results cannot be generalized to the Somali country because those living in the United States have better health-care facilities than those living in Somalia, which has poor health-care facilities, economic problems, and political issues [Tahtabasi, et al., 2020].

While there are no definitive prevalence and incidence studies for Somalia, some estimates indicate that the incidence of breast cancer accounts for about (1 892) new (29.5%) of all cancers (6 411) in 2020, while Cervix uteri cancer, Colorectum cancer, Ovary cancer and Thyroid cancer represent about 1 055 new cases (16.5%), 358 new cases (7.3%), 307 new cases (5.6%) and 231 new cases (3.6%) of all cancers respectively in 2020 [<https://gco.iarc.fr/today>, Globocan, 2020]. These are the top five cancers in Somalia in terms of prevalence, with the crude ratio for each cancer as follows: Breast cancer 23.7%, Cervix uteri cancer 13.2%, Colorectum cancer 4.5%, Ovary cancer 3.9% and Thyroid cancer 2.9% per 100,000 women in a 5-year cohort study [<https://gco.iarc.fr/today>, Globocan, 2020].

According to a 5-year cohort study of all cancers, the prevalence of breast cancer in both sexes is around (2472) cases (18.0%), Cervix uteri cancer (1320) cases (10%), Colorectum

cancer 840 cases (6.4%), Leukemia cancer 575 cases (5.7%), and NHL cancer 646 cases (4.9%) [<https://gco.iarc.fr/today>, Globocan, 2020].

The gap was that there were no articles discussing breast cancer for women; instead, all reviewed articles or journals focused on breast cancer for both sexes. According to Somalia, since the civil war, no government organizations were dealing with the cancer sector for a long time, but some private hospitals were opened after 2010, so the statistical journals talking about breast cancer were not more and they were writing for general cancer but not for only breast cancer. That is why I am emphasizing to talk about breast cancer in women in Somalia, because of the data collected from Somalia, most participants were women only.

When well-structured, mathematical models can aid in this comprehension by providing a more in-depth look at some of the properties of the cancer survival distributional pattern in Africa. Survival analysis is the time when an event occurs, which means from the time to begin following up on the event until the event of interest is censored or occurs, a prospective cohort study is the most common method used. The study of patterns of event times, the evaluation of distributions of survival times in various kinds of people, and the review of whether and how much certain variables influence the probability of an event of interest are all goals of survival analysis [Kartsonaki, et al., 2016].

The Kaplan-Meier estimator and proportional hazards models, for instance, Cox's are two survival analysis methods that are widely used to study survival time, first enables the estimation of survival functions, while the second enables the evaluation of explanatory variables on the hazard ratio. It has the same event of interest, such as death or survival, at the same time [Ferraz, et al., 2017].

The main goal is to establish or calculate the survival rates, prevalence and to evaluate independent variables that cause breast cancer and other factors that can affect breast cancer for women only in Mogadishu, Somalia. Developing a model using breast cancer data in Mogadishu, Somalia, the appropriate model will be the proportional hazards model. The model was useful for this research because it allowed researchers to exclude covariates that appeared to not affect survival rate, resulting in a survival model with variables that had a significant impact on breast cancer survival that allowed researchers to discuss the association between output and independent variables, whether significant or not, and step-by-step building tables involved the survival.

1.2 Statement of Problem

Breast cancer affects both sexes and is the most common cancer in the world, according to other cancers that have harmed or killed humans in recent years. <https://gco.iarc.fr/today>. Female breast cancer is the most common cancer in Somalia, and it is also the leading cause of death among Somali women, according to new cases in 2020 <https://gco.iarc.fr/today>.

Somalia is located in east Africa and has been ravaged by civil war, which has hurt the health and economic sectors, because most people are poor and health care is limited, more women develop breast cancer, however, there is no national bureau of statistics that deals with cancer and collects data, nor is there any other cancer registry office that can register patients, the prevalence and deaths of female breast cancer in Somalia was increasing last year [Tahtabasi, et al., 2020]. For these reasons, the main goal of this research is to determine the survival rates and risk factors for breast cancer in Mogadishu, Somalia. To analyze independent variables with time independence, the appropriate model is Cox's proportional hazard model, which also employs Kaplan-Meier and long-rank methods.

RESEARCH QUESTION: What are the advantages of assessing breast cancer risk factors for Somali women?

NULL HYPOTHESIS: Breast cancer risk variables were assessed in a research of Somali women and were found to have a statistically significant impact on the woman.

1.3 Objectives

General Objective: The main goal is to determine the survival rates and risk factors for breast cancer in Mogadishu, Somalia.

Specific Objective:

- I) Evaluating the prevalence of breast cancer for women patients in Mogadishu, Somalia.
- II) Examining the survival and the risk factors for breast cancer for women in Mogadishu, Somalia.

1.4 Justification/Significance of Study

Cancer data can only be obtained from special hospitals, which are no longer in operation; the cancer registry center is not yet operational in Somalia; and the journals that were discussing or publishing analysis and information about cancer are no longer in operation, which is why I am going to talk about cancer in Somalia, specifically women's breast cancer, in the hope that it will assist the government and residents in learning the facts and raising awareness about the disease.

This research is very important among patients who suffer from breast cancer, inspiring us to keep up a balanced lifestyle because of getting information about survival rates, KM, and other tables that help us to predict a way of living for patients. Findings of the research can be used to help ordinary women, as well as governments and other agents like non-government organizations, gain a better understanding of the risk factors for breast cancer. This research could lead to better healthcare planning in Mogadishu,

Somalia. The study's target aligns with the World Cancer Report's request for early diagnosis, treatment, hospice care since it is affecting older and palliative services.

1.5 Structure of the thesis

I intended to do the following in this research: chapter one is the introduction. It has five subsections. They are Background, Statement of Problem, Objectives, Justification/Significance of Study, and Structure of the thesis. Chapter two is the Literature review. It contains Breast Cancer, Survival analysis, and an Overview of the Literature Review. The third chapter is a research method, and it includes the sections listed below. Research Design, Population of the Study, Sample, Procedure, Survival analysis, Kaplan Meier Estimator, Log-Rank Test, Cox Proportional Hazard. Chapter four is data analysis and results, holding Introduction, Descriptive statistics, Survival analysis. And finally, chapter five contains the Summary, Conclusion, Recommendations, and Further Research.

2 Literature Review

2.1 Breast Cancer

In 2012, approximately 14.1 million new cancer cases were confirmed, with nearly 8.2 million People dying from cancer all over the world. Breast cancer is the most commonly diagnosed cancer in women and the leading cause of mortality in the world, with approximately 1.7 million new infections and approximately 522 thousand related deaths [Lukong, et al., 2017].

The highest frequency of diagnosed cancer in women is breast cancer in developing and developed nations. Breast cancer is the most prevalent cancer among Iranian women, accounting for 21.4 percent of all malignancies, according to a study published by the Iranian Center for Disease Prevention and Control, Ministry of Health and Medical Education, in 2000. Breast cancer is estimated to affect 8% to 10% of women in Europe and the United States, and Asian nations, on the other hand, have the smallest incidence, at around 1% [Rezaianzadeh, et al., 2009].

Breast cancer penitents in Africa countries begin by identifying a significant number of patients with the progressed disease and also have minimal access to cancer education, testing, and treatment; according to a study conducted in North Africa, Egyptian and Tunisian women have a risk side view that is more supportive of developed countries. This includes a higher average number of infants, a younger average age at first birth, a longer average time of breastfeeding, a younger average age at menopause, lower average age at menopause, a low rate of contraception use, and a low average of alcohol consumption, in Africa, most patients have advanced-stage breast cancer, like 89.6% in Kenya and 72.8% in Nigeria of breast cancer patients having advanced-stage disease, the incidence of advanced-stage breast cancer was stated to be between 50% and 55% in South African research, a Stage III and IV breast cancers have a 33% incidence from research in Moroccan countries, in contrast to Africa, breast cancer presents to patients in developed countries at a younger age. Women with cancer in West Africa are on average 35 to 45 years old, which is 10 to 15 years younger than women in developed countries [Vanderpuye, et al., 2017].

In Somalia, the reality of cancer incidence is unknown. Cancer incidence is most determined by data collected from hospitals in the capital or urban areas, since 32 countries, such as Somalia, are unable to collect local data, the global cancer statistics figures suggest cancer incidence forecasts based on the average rates of surrounding countries, there is no local data. Assessments of cancer incidence in Somalia are calculated as the average of Ethiopia and Kenya, as a consequence, global cancer statistics data vary from the find-

ings of our research, because Somalia lacks a national cancer registry, the exact findings are uncertain, as they are in other countries [Tahtabasi, et al., 2020].

The prevalence rate for breast cancer in Somalia in 2020 was 79.7 per 100 000 people, or 0.0797% of the population; the incidence rate for breast cancer in women of all ages in Somalia was estimated at 23.7 per 100 000 women for age-standardized in 2020; and the death rate for breast cancer in women of all ages in Somalia was 14.9 per 100 000 women for age-standardized in 2020 [[https : //gco.iarc.fr/](https://gco.iarc.fr/) today globacan].

Many factors, including sex, aging, estrogen, family history, gene mutations, an unhealthy lifestyle, obesity, age at menarche, age at first life birth, stages, whether the medical facility used is chemotherapy or not, and finally, tumor size and grades, have been linked to an increased risk of developing breast cancer. Stated that from [Kantelhardt, et al., 2015, Vanderpuye, et al., 2017].

Tanzanian researchers matched 115 breast cancer patients with 230 controls in another case control analysis, Premenopausal and postmenopausal breast cancer risk grew with a higher estimated BMI at age 20 (OR = 1.31, 95% CI 1.11 - 1.55), and also Premenopausal breast cancer risk was decreased by later menarche and long-term lactation (OR menarche = 0.74, 95% CI 0.56 - 1.00; OR lactation = 0.98, 95% CI 0.97 - 0.99) [Kantelhardt, et al., 2015].

2.2 Survival analysis

Some health trials are attempting to classify the estimates of survival or any statistical inference that involves patients. Many scientists believe that examining survival data requires the use of three traditional statistical methods: parametric when the data is dependent on statistical distributions, semiparametric when the distribution is indeterminate or the model has a finite-dimensional component (easy to research and understand), and non-parametric when the data is not dependent on any distribution or there are no assumptions [Abadi, et al., 2014, Kartsonaki, et al., 2016].

Survival analysis is a type of statistical study that is highly precise, so the objectives of survival analysis are to look at the time until the event of interest occurs, and then failing time or survival time is another word for such a duration of concern, so the duration of a survival time can be calculated using days, months, weeks, years, or other scales, the Kaplan Meier estimation is the most statistical tool for evaluating data and comparing two different people classes. Considering the log-rank in survival research, the KM calculation is among the most effective statistical techniques for determining the likelihood of a patient surviving for that same amount of time following medication [Etikan, et al., 2017]. Despite many of the problems involved with subjects or conditions, the KM calculation is the easiest method for estimating survival across periods. In the Kaplan-Meier calculation, curves are employed to calculate the occurrences, censorship, and survival likelihood [Etikan, et al., 2017, Dakhil, et al., 2012].

The research population is made up of a retrospective cohort collected from inhabitants in the municipality of Campinas, Sao Paulo from the Cancer Registry between January 1st, 1993 and December 31st, 2011, followed up for 18 years, and the research is built on the records of 524 women who have been diagnosed with breast cancer, the overall survival rate calculated using the Kaplan-Meier procedure (K-M) was 60.8 percent, implying that women in the postmenopausal era had a survival rate of 76.3% after five years, with 62% and 68% of these women divided by aspect of service usage. [Ferraz, et al., 2017].

In the health sciences, the Cox proportional hazard analysis is the very commonly used semi-parametric, as opposed to parametric, the Cox model has less presumption, which is one of the causes of its popularity, the Proportional Hazards (PH) models presume that perhaps the risk ratio or odd ratio of two individuals is unaffected by the passage of time, Just 5% of all research that uses the Cox PH model regarded the underpinning principle, according to a study of survival data in cancer articles, other changes or models have to be applied for survival data analysis if the proportional expectation is not met, schoenfeld residuals were employed to verify the PH assumption, the PH presumption was tested utilizing Schoenfeld residuals, the Cox PH model was employed breast cancer patients in r in Southern Iran total of 15830 sick persons were included in the research, the average age of the patients became (59.1 plus or minus 13.4), while around 30.4% from the data was under the age of 50, while 69.6% was over the age of 50 but 58.9% of them was censored from the report, there is mortality from breast cancer or other causes around 41.1% of patients in this research, participants who underwent chemotherapy had such a greater risk (HR =3, CI: 2.29 - 3.93) than those who did not, and the Chemotherapy was linked to a lower survival rate in diagnosed with first stage of breast cancer [Abadi, et al., 2014].

2.3 Overview of Literature Review

Overall, the research listed earlier, discovered the same prediction factors linked to the growth of breast cancer. Factors to consider include sex, aging, an unhealthy lifestyle, obesity, age at menarche, age at first life birth, tumor stages, tumor size, tumor grades, and chemotherapy. According to the research, becoming aware of the signs and detecting them early by testing can lead to an earlier diagnosis and better care outcomes. Literature concluded that awareness of the symptoms, and early detection through screening can help lead to earlier diagnosis, resulting in improved treatment outcomes.

Survival analysis is a collection of methods for comparing the chances of the danger of death or any other occurrence related to various medications or classes, in which the danger evolves with time. In this chapter, only the most widely used survival types are discussed, such as the Kaplan-Meier approach is used to estimate the survival curve, the log-rank methods were used to compare two classes whether they differ or same, and the Cox's proportional hazards approach affords for the use of other explanatory variables [Bewick, et al., 2004].

3 Research Methods

3.1 Research Design

Because the research data was secondary data, the analysis approach used in this study was a quantitative and qualitative method with a retrospective cohort study since the research used quantitative and qualitative data sets that came from research problems. This research was carried out in Mogadishu, Somalia.

3.2 Population of the Study

The breast cancer patients who had visited the clinic were the target population. The data was received from Osman hospital, especially the cancer registry office for that hospital in Mogadishu, Somalia from February 2015 up to March 2020. A total of 551 cancer cases in Osman hospital in Mogadishu, Somalia was registered. All of them were women because there were no men who went to the hospital in Mogadishu, Somalia to screen or check whether they had breast cancer or not. The data from March 2020 until 2021 was not ready to use for this study since it can't be obtainable during this study.

3.3 Sample

The sample for this study was made up of all patients diagnosed with breast cancer in Mogadishu, Somalia. The participants in this research were all women patients in Mogadishu, Somalia who had been diagnosed with breast cancer. There were 551 cancer cases registered in the registry, according to the data collected from Osman hospital in Mogadishu, Somalia. All patients diagnosed with breast cancer were chosen to be in the sample study. A simple random sampling technique was applied to select subjects from Osman hospital in Mogadishu, Somalia. This approach guaranteed that the sample collected was representative of the entire population and that every subject had an equal probability of being included in the sample.

3.4 Procedure

The research focused on women breast cancer patients under the age of 80. People had visited the breast cancer office at Osman Hospital, which offered testing or screening at all stages of cancer. Before doing any data analysis, the data had to be cleaned. This included missions such as removing duplicate cases and factors, choosing cases and factors, removing the patient's identity card number, matching files, as well as other data preparation activities.

Variables in the breast cancer data collected for the research were age in years, age at first life birth in years, tumor size in millimeters, tumor grade ordered for three groups (one as I, two as II, and three as III), number of positive nodes, Breast if left or right

possession, Obesity for two groups if obese or not, and Chemotherapy for breast cancer for two groups if they use Chemotherapy drugs to destroy breast cancer cells, or if they use other treatments, such as surgery, radiation or hormone therapy, The time variable for survival is the difference between the beginning follow up time and the endpoint or the last time for patient contact, so at that time the patient either died or survived.

3.5 Data analysis

The data gathered was coded and recorded into excel. After that, the data was imported into R, which was used to analyze this research. The development of descriptive and inferential statistics was part of the data analysis process. Tables and graphs were used to display the descriptive statistics. The data were subjected to survival analysis methods. The Kaplan-Meier curves and log-rank tests were among the inferential statistics produced. The variations in survival across various covariates and their outcomes were represented using Kaplan-Meier curves.

The Cox Proportional Hazards model was used to examine the connection between breast cancer patients' survival time and the risk factors.

The following is the model formula:

$$Cox(x) = h(X) = h_0(t)(\exp(\beta_1 \text{chemotherapy} + \beta_2 \text{Breast} + \beta_3 \text{Obesity} + \beta_4 \text{tgrade} + \beta_5 \text{age} + \beta_6 \text{age.fbirth} + \beta_7 \text{tsize} + \beta_8 \text{pnodes}))$$

Where $h(t)$ denotes the expected risk at time t . $h_0(t)$ is the baseline hazard function. The log-baseline hazard is represented by the constant in this model, since all independent variables or risk factors are equal to zero

3.6 Survival analysis

Survival analysis is a set of statistical techniques for data processing where the dependent variable is the time before an occurrence happens. The time variable is sometimes referred to as "survival time" in survival analysis because it indicates how long a person has "survived" in a given era of time. Since the occurrence of concern is normally death, or any negative individual encounter, we named failure time [Kleinbaum, et al., 2010].

Statistics that show the length of time between a certain starting point and the occurrence of a specific event called Survival periods, such as time between the start and finish of a recovery period or the time between a disease diagnosis and death. [Bewick, et al., 2004].

3.6.1 Censoring

Censoring is a central methodological issue that must be considered in most survival studies. Censoring happens if we have sufficient knowledge of a person's survival period but not at the right time. [Bewick, et al., 2004].

From this journal [Bewick, et al., 2004], Censorship could happen for one of three motives:

1. An individual survives until the follow-up time is finished;
2. During the observation time, an individual is gone out or lost for the follow-up;
3. A participant leaves from the research since he or she died of another problem or any other cause

The research subjects are studied not only when the case occurs, but also when they are followed until they are removed from the research or until the follow-up time is completed, also we don't know what happens after the follow-up time is completed, so that is named censored, which means we don't know the exact time the event occurred for an included observation, censoring must be a non-informative process with no significant impact on survival performance, censoring can occur at any time during a report, whether at the start, closing time, or any other stage, it can be right censoring if the patient left after the set time or completed the research, but we don't know how the patients felt afterward, or it can be left censoring if a person is included in a study even though the occurrence of concern occurred before recruitment [Etikan et al., 2018, Kartsonaki, et al., 2016].

3.6.2 Survival Function: [S(t)]

From these books [Bewick, et al., 2020, Kleinbaum, et al., 2010] we found that the probability of an object defined by T surviving past time t is given by the survival function:

1. The period is described as $t \in \infty$.
2. S(t) does not increase, i.e., $S(t_1) \geq S(t_2)$ for $t_1 \leq t_2$
3. The likelihood or chance of living at time zero is one. $t = 0, S(t = 0) = 1$, i.e., also $(S(t) = 0$ for $t \rightarrow \infty)$

Since S(t) is a likelihood:

$$S(t) = \text{Prob}(T > t) \implies S(t) = 1 - F(t) \quad (1)$$

Where S(t) denotes the survival function and F(t) denotes cumulative distribution function of T. As a result from question (1) there is a probability density f that has the characteristic

$$S(t) = \int_t^{\infty} f(\tau) d\tau \quad (2)$$

This means that we can get $S(t)$ by differentiating it:

$$f(t) = -\frac{d(S(t))}{dt} \quad (3)$$

T's expected value, or mean, is received by

$$\mu = \mathbb{E}(T) = \int_0^{\infty} t f(t) dt \quad (4)$$

As expressed by f , this is the patients' average life expectancy. Using Equation (3) and integrating by parts, it can be shown that the survival function $S(t)$ can be used to calculate the average life expectancy:

$$\int_a^b v du = uv|_a^b - \int_a^b u dv \quad \text{this formula called integration by parts}$$

$$\mu = \mathbb{E}(T) = \int_0^{\infty} t f(t) dt = - \int_0^{\infty} t \frac{d(S(t))}{dt} dt$$

where $v = t$ there fore $dv = dt$ where $du = d(S(t))$ there fore $u = S(t)$

After that we make substitution

$$- \int_0^{\infty} t \frac{d(S(t))}{dt} dt = - \left[(tS(t))|_0^{\infty} - \int_0^{\infty} S(t) dt \right] =$$

According to properties we get $t = 0, S(t = 0) = 1$ and $(S(t) = 0$ for $t \rightarrow \infty)$ There fore

$$- [(\infty) * S(\infty) - (0) * S(0)] = - [(\infty) * (0)] = 0 \text{ there fore } - [(tS(t))|_0^{\infty}] = 0$$

Just remaining is part of the integration

$$- \int_0^{\infty} t \frac{d(S(t))}{dt} dt = - \left[- \int_0^{\infty} S(t) dt \right] = \int_0^{\infty} S(t) dt$$

Finally we obtain the mean life expectation is

$$\mu = \mathbb{E}(T) = \int_0^{\infty} S(t) dt \quad (5)$$

3.7 Kaplan Meier Estimator

While evaluating an approximate survival curve, it's important to keep in mind that risk collection shouldn't be too minimal. Furthermore, if a procedure is intended to postpone the beginning of an incident, care should be taken to ensure that sufficient follow-up time is provided. That is since the risk set is limited and the majority of study subjects become censored, extrapolating the calculated survival likelihood and relevant CI well past the last reported event period should be done with caution [Carter, et al., 2009].

Free of assuming a fundamental probability distribution, the Kaplan-Meier approach can be used to approximate the curve from observed survival times [Bewick, et al., 2004]. The event of interest marks the end of the sequential time of defined survival; in K-M analysis, this was defined as an interval and plotted as a horizon line [Rich, et al., 2010]. The Kaplan-Meier method is the most popular method used for survival analysis. Together with the log-rank test, it may provide us with an opportunity to estimate survival probabilities and to compare survival between groups [Jager, K, et al., 2008].

Each topic is described by three aspects when performing a K-M survival study [Rich, et al., 2010]:

- I) The sequence of their times
- II) Their current situation at the termination of their serial period, whether the event happens or censored (alive).
- III) And the research community they're a part of.

Further, When constructing survival time likelihood and curves, the sequence times of the person subjects are organized from smallest to greatest, regardless of where they joined the sample, this strategy ensures that all members of the community start studies at the same time and that they all survive before one of them dies [Rich, et al., 2010].

The Kaplan-Meier survival curve is the likelihood of living for a specified period as time is divided into several small intervals. In general, we check three presumptions in the observations which we need to analyze [Goel, et al., 2010].

1. At any point in time, we believe that censored persons have an equal chance of survival as those who are observed.
2. We conclude that the survival rates for participants enter the early and late in the research are equal.
3. We presume that the occurrence takes place at the required time.

This can be problematic in some cases where the occurrence will be observed during a routine inspection. We know that the incident took place in the middle of two exams. Participants could be followed up on more often and at shorter periods to provide a more reasonable measure of survival; The "product limit estimate" is another name for the Kaplan-Meier method [Goel, et al., 2010].

3.7.1 Kaplan-Meier Estimator for the Survival Function

From these two books [Bewick, et al., 2020, Kleinbaum, et al., 2010] quoted that the following formulas.

The question below for the Kaplan-Meier rate for failing time is called the general formula.

$$S(t_{(f)}) = S(t_{(f-1)}) * \hat{P}_r (T > t_{(f)} | T \geq t_{(f)}) \quad (6)$$

Kaplan Meier = product limit

$$S(t_{(f-1)}) = \prod_{i=1}^{f-1} \hat{P}_r (T > t_{(i)} | T \geq t_{(i)}) \quad (7)$$

Let

1. N = the sample size that is being investigated.
2. m = the total number of fatalities
3. $t_1 < t_2 < \dots < t_k$ are the occasions when deaths were noticed in a certain sequence.
4. k = The number of different occasions the incident has happened.
5. d_j = the total number of people who died at a particular point in time t_j for $1 \leq j \leq k, d_1 + d_2 + \dots + d_k = m$
6. c_j = the number of people whose lives have been censored from t_j to t_{j+1} ($1 \leq j \leq k$)
 $c_0 + c_1 + c_2 + \dots + c_k = N - m$
7. n_j = the number of people whose lives were put in jeopardy just as time was running out t_j .

The Kaplan-Meier method is calculated as follows:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) \quad (8)$$

$$Var[\hat{S}(t)] = ([S(t)]^2) * \left[\sum_{t_j \leq t} \left(\frac{d_j}{n_j - d_j} \right) \right] \quad (9)$$

Confidence Interval $(1 - \alpha) * 100\%$ for $S(t)$ CI:

$$CI = \hat{S}(t) \pm \left(z_{\frac{\alpha}{2}} \right) * \left(\sqrt{Var(S(t))} \right) \quad (10)$$

From this book [Bewick, et al., 2020], the Kaplan Meier table (1) on the next page tells us how to calculate the table columns.

Sample	size	N					
j	t_j	d_j	c_j	n_j	$\frac{d_j}{n_j}$	$1 - \frac{d_j}{n_j}$	$S(t) = \prod_{j \leq t} (1 - \frac{d_j}{n_j})$
0	$t_0 = 0$	$d_0 = 0$	c_0	$n_0 = N$			1
1	t_1	d_1	c_1	$n_1 = N$			$1 \cdot (1 - \frac{d_1}{n_1})$
2	t_2	d_2	c_2	n_2			$1 \cdot (1 - \frac{d_1}{n_1}) \cdot (1 - \frac{d_2}{n_2})$
3	t_3	d_3	c_3	n_3			$1 \cdot (1 - \frac{d_1}{n_1}) \cdot (1 - \frac{d_2}{n_2}) \cdot (1 - \frac{d_3}{n_3})$
\vdots	\vdots	\vdots	\vdots	\vdots			\vdots
k	t_k	d_k	c_k	n_k			$1 \cdot (1 - \frac{d_1}{n_1}) \cdot (1 - \frac{d_2}{n_2}) \dots (1 - \frac{d_k}{n_k})$
Total		m	$N - m$				

Figure 1. A visualization for the meaning of Kaplan-Meier table

3.8 Log-Rank Test

A mathematical hypothesis test known as the log-rank method can be used to compare two survival curves. It is applied to evaluate the null hypothesis that the group survival curves are similar, the log-rank method adds up the χ^2 to every event time on every category. To match the complete curves of every category, the aggregate outcomes at each group are added to get the final χ^2 [Rich, et al., 2010].

According to these [Kleinbaum, et al., 2010, Bustan, et al., 2018], i summarized that we presume we have two survival curves that refer to two categories of cases, for example, one category was given medicine and another was given a dummy. A hypothesis test, as well as the null and alternative, can be used to make a statistical comparison.

The following is a possible hypothesis:

Hypothesis 1 In terms of survival, there is no distinction between the two categories.

Hypothesis 2 In terms of survival, there is the distinction between the two categories.

Expected value one for long rank:

$$e_{1t} = \left(\frac{n_{1t}}{n_{1t} + n_{2t}} \right) * (m_{1t} + m_{2t}) \quad (11)$$

Expected value two for long rank:

$$e_{2t} = \left(\frac{n_{2t}}{n_{1t} + n_{2t}} \right) * (m_{1t} + m_{2t}) \quad (12)$$

from question (11) and (12) We can write likelihood for two groups in a different way, as follows.

$$e_{2t} = (P_r(\text{group 1}) * (m_{1t} + m_{2t})) \quad (13)$$

$$e_{2t} = (P_r(\text{group 2}) * (m_{1t} + m_{2t})) \quad (14)$$

Where (m_{1t}) represents the number of participants in category one who failed at a certain time, and (n_{1t}) Is the risk set by category one at the time. And (m_{2t}) represents the

number of participants in category two who failed at a certain time and (n_{2t}) is the risk set by category one at the time.

As a result, the question (15) in log rank statistics is a one-degree-of-freedom chi-square distribution.

$$\text{Log-rank statistics} = \sum_{i=1}^k \left(\frac{(O_i - E_i)^2}{E_i} \right), \text{ where } k \text{ is a group numbers} \quad (15)$$

3.9 Cox Proportional Hazard

Kaplan Meier and Log-rank tests are used for uni-variate analysis. According to the multivariate analysis, the Cox Proportional Hazard (CPH) models are used.

First, when we are talking about Cox Proportional Hazard models, that means we are using the Hazard function, so I want to talk about the Hazard function.

3.9.1 Hazard Function, $h(t)$

From this book [Kleinbaum, et al., 2010] Provided that the person lived up to time t , the hazard function $h(t)$ offers the instantaneous probability of each unit of time for the incident to happen. The hazard function, in comparison to the survivor function, which emphasizes not failure, reflects on having failed, or the accident happening. Thus, the hazard function could be thought of as providing the inverse of the knowledge provided by the survivor function.

[Bewick, et al., 2020, Kleinbaum, et al., 2010], the hazard function may rise, fall, stay the same, or imply a more difficult system. It's named a birth tub hazard function when the hazard function has been both declining and growing.

This journal [Bustan, et al., 2018, Kleinbaum, et al., 2010], i summarized the hazard function $h(t)$ has the below properties:

- $h(t) \geq 0$ for all t ,
- $h(t)$ there is no upper limit,
- $h(t)$ can take on any shape.

The meaning of $h(t) = 0$ is that no event happened in Δt . The cumulative hazard function describes the accumulated risk up to time t given by

$h(t) = 0$ denotes that no occurrence occurred in time Δt . The cumulative hazard function defines the risk that has increased up to time t that is given by:

$$H(t) = \int_0^t h(\tau) d\tau \quad (16)$$

$$f(t) = \lim_{\Delta t \rightarrow 0} \left(\frac{P(t \leq T < t + \Delta t)}{\Delta t} \right) \quad (17)$$

$$h(t) = \lim_{\Delta t \rightarrow 0} \left(\frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right)$$

$$h(t) = \lim_{\Delta t \rightarrow 0} \left(\frac{P(t \leq T < t + \Delta t)}{\Delta t} \right) * \left(\frac{1}{P(T \geq t)} \right)$$

$$h(t) = \lim_{\Delta t \rightarrow 0} \left(\frac{P(t \leq T < t + \Delta t)}{\Delta t} \right) * \left(\frac{1}{S(t)} \right)$$

$$h(t) = \left(\frac{f(t)}{S(t)} \right) \quad (18)$$

Finally the relationship between or the connection between $h(t)$ (or $H(t)$) and $S(t)$ is given by :

$$S(t) = \text{Exp} \left[- \int_0^t h(t) du \right] \quad (19)$$

$$h(t) = - \left[\left(\frac{1}{S(t)} \right) * \left(\frac{d(S(t))}{dt} \right) \right] \quad (20)$$

3.9.2 Cox Proportional Hazard Model

The log-rank method is applied to see if there is a disparity in survival times among classes, but it excludes many predictors. An appropriate model for multiple variables in survival is cox's proportional hazards model, it allows the disparity in survival times of different patient populations to be evaluated when other variables are taken into account. Cox's model makes no assumptions regarding the hazard's likelihood distribution [Bewick, et al., 2004].

The hazard ratio is assumed to be steady, which is the key presumption, its calculated number is referred to as relative risk [Kartsonaki, et al., 2016].

To check the CPH presumption could be evaluated using Schoenfeld residuals, Stratification is used to modify predictors that do not fulfill the PH assumption, while adding in the model is used to adjust predictors that are doing [Abadi, et al., 2014].

To determine the CPHM, which is a semi-parametric regression model the hazard function is given by:

$$h(t, X) = h_0(t) * (\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + x_p)) \quad (21)$$

Or according to the hazard ratio, the Cox proportional hazards regression model is written the following.

$$\text{hazard ratio} = \frac{h(t, X)}{h_0(t)} = \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p) \quad (22)$$

Schoenfeld residuals was used to test the PH assumption

Schoenfeld residuals are calculated for any subject who experiences an occurrence of any independent variables inside this model. The statistical test assumes that if the PH presumption is maintained for a given variable, the Schoenfeld residuals for this kind of variable are unrelated to survival time [Kleinbaum, et al., 2010].

From this book [Tableman, et al., 2003], i summarized the following formulas the r_{sjk} denotes the k_{th} Schoenfeld residual specified for the k_{th} subject on the j_{th} independent covariate $x^{(j)}$.

$$r_{sjk} = \delta_k \left\{ \left(x_k^{(j)} \right) - \left(a_k^{(j)} \right) \right\} \quad (23)$$

$x_k^{(j)}$ represents the amount of the j_{th} independent covariate on the k_{th} person in the sample, whereas δ_k represents the censoring measure for the k_{th} object.

$$\left\{ \left(a_k^{(j)} \right) = \frac{\left(\sum_{m \in \mathcal{R}_{(yk)}} \exp(\underline{x}'_m \hat{\underline{\beta}}) x_m^{(j)} \right)}{\left(\sum_{m \in \mathcal{R}_{(yk)}} \exp(\underline{x}'_m \hat{\underline{\beta}}) \right)} \right\}$$

$(\mathcal{R}_{(yk)})$ denotes the risk set at time yk , and $\left(a_k^{(j)} \right)$ denotes the weighted mean of independent covariates values across persons at risk at time yk .

dfbetas to determine the impact of every observation

We need to see how any observation affects the estimation of the $\hat{\underline{\beta}}$ from the $\underline{\beta}$. We look at which variable elements are present and $\left(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(k)} \right)$ have excessively high absolute values. Repeat this procedure for any of the n samples. This test appears to be identical to dfbetas in linear models [Tableman, et al., 2003].

The k th dfbeta has the following definition:

$$\text{dfbeta}_k = I(\hat{\underline{\beta}})^{-1} (r_{s1k}^*, \dots, r_{smk}^*)' \quad (24)$$

where $I(\hat{\underline{\beta}})^{-1}$ is the Fisher matrix's inverse.

$$r_{s,jk} = \delta_k \left\{ \left(x_k^{(j)} \right) - \left(a_k^{(j)} \right) \right\} - \exp(\underline{x}'_k \hat{\underline{\beta}}) \left(\sum_{t_i \geq yk} \frac{\left\{ \left(x_k^{(j)} \right) - \left(a_k^{(j)} \right) \right\}}{\sum_{l \in \mathcal{R}(t_i)} \exp(\underline{x}'_l \hat{\underline{\beta}})} \right) \quad (25)$$

The above formula is Schoenfeld residual minus the effect across all risk sets that include the k_{th} object.

The appropriate functional form of the continuous variable is determined using Martingale residuals

From this book [Tableman, et al., 2003] i summarized , the deviance residuals are calculated as follows:

$$R_{Di} = \text{sign}(R_{Mi}) (2[-R_{Mi} - \delta_i \log(\delta_i - R_{Mi})])^{\frac{1}{2}} \quad (26)$$

$\text{sign}(R_{Mi})$ denotes the sign function, If the statement is positive the value will be 1, 0 if the value is zero If the value is negative, the result is (-1). R denotes the i th parson's martingale residual.

$$R_{Mi} = \delta_i - R_i, \quad i = 1, 2, 3, \dots, n \quad (27)$$

With a mean of zero, the martingale residuals have a skewed distribution.

4 Data Analysis And Results

4.1 Introduction

4.1.1 Check for violation of proportional hazard

There are some assumptions in the Cox proportional hazards model to determine whether a Cox model that has been fitted fully explains the results or not.

- 1) The Proportional Hazard model assumes that the hazard ratio of two patients remains constant over time. To check this assumption, we will employ Schoenfeld residuals. Then we will check the p-value whether it is less than 0.05 or not. Therefore, any finding with a p-value less than 0.05 indicates that the proportionality assumption has been broken.
- 2) To determine the outlier for the independent variables, we will use the Deviance residual or the dfbeta values.
- 3) For continuous independent variables, the Proportional Hazard assumption is log-linear. That means the continuous explanatory variables are assumed to have a linear shape. To check this assumption, we will use Martingale residuals, then we will check which one fits the data very well.

So, we begin to check these three assumptions step by step using R studio software, graphs, and tables. First as the result from the graph (4.1) below, for each variable was not statistically significant, the global test is not significant. That means we can use a cox proportional model that is independent of time.

```
> sursom=read.csv("F:\\breast cancer Somalia.csv")
> names(sursom)
[1] "chemotherapy" "age" "age.fbirth" "tsize" "tgrade"
[6] "pnodes" "Breast" "time" "Obesity" "Censoring"
> cox.rg=coxph(Surv(time, Censoring)~chemotherapy+tsize+Breast+age+tgrade+age.fbirth+
+ tsize+Obesity+tgrade+pnodes, data = sursom)
> cox.tst=cox.zph(cox.rg)
> cox.tst
      chisq df  p
chemotherapy 1.88e-05 1 1.00
tsize        6.64e-01 1 0.41
Breast       2.52e+00 1 0.11
age          2.50e-01 1 0.62
tgrade       4.54e+00 2 0.10
age.fbirth   2.14e+00 1 0.14
obesity      1.47e-01 1 0.70
pnodes       4.64e-01 1 0.50
GLOBAL      1.14e+01 9 0.25
> cox.rg
Call:
coxph(formula = Surv(time, Censoring) ~ chemotherapy + tsize +
Breast + age + tgrade + age.fbirth + tsize + obesity + tgrade +
pnodes, data = sursom)

      coef exp(coef) se(coef)      z      p
chemotherapy -0.322080  0.724640  0.117726 -2.736 0.006222
tsize         0.017634  1.017791  0.005275  3.343 0.000828
BreastRight   0.339324  1.403998  0.115086  2.948 0.003194
age           0.016042  1.016171  0.006831  2.348 0.018856
tgradeII     0.543875  1.722670  0.169713  3.205 0.001352
tgradeIII    0.581966  1.789553  0.217839  2.672 0.007550
age.fbirth   0.045611  1.046667  0.021859  2.087 0.036920
obesityobes  0.246820  1.279949  0.115287  2.141 0.032280
pnodes       -0.024098  0.976190  0.010403 -2.316 0.020535

Likelihood ratio test=101.6 on 9 df, p< 2.2e-16
```

Figure 2. Cox Proportional Hazard assumptions for breast cancer data

According to this graph (3) below there is no trend or pattern with time-based on the graphical analysis.

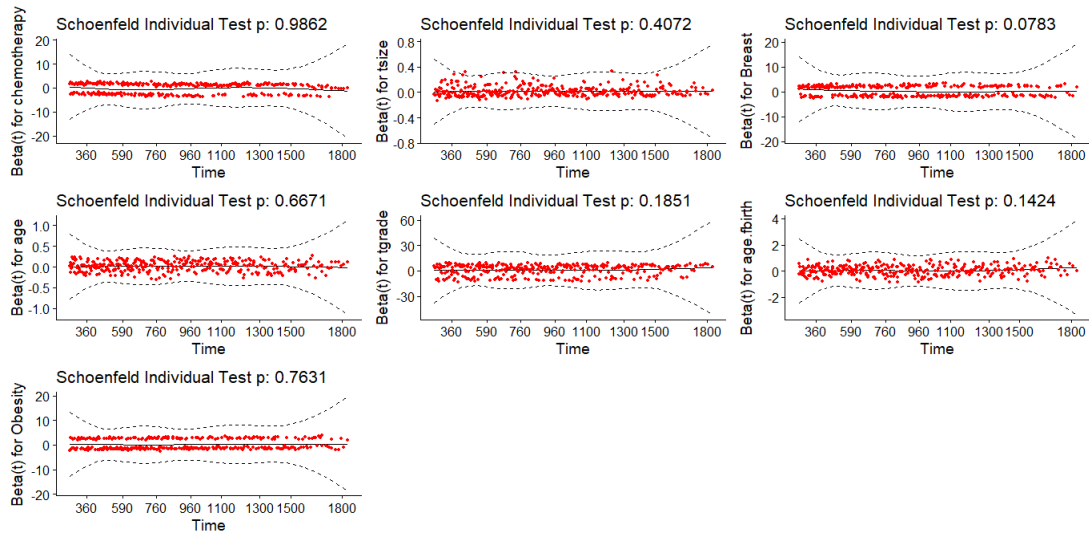


Figure 3. The use of graphs to check for time independence shows us whether the variable is batten tome or not

According to the number of the largest dfbeta values from the plot in the graph (4) below, none of the findings are highly influential, even though some of them have a large value when compared to the age at first birth and tumor size for women with breast cancer, indicating that there is no outlier in our model or data.

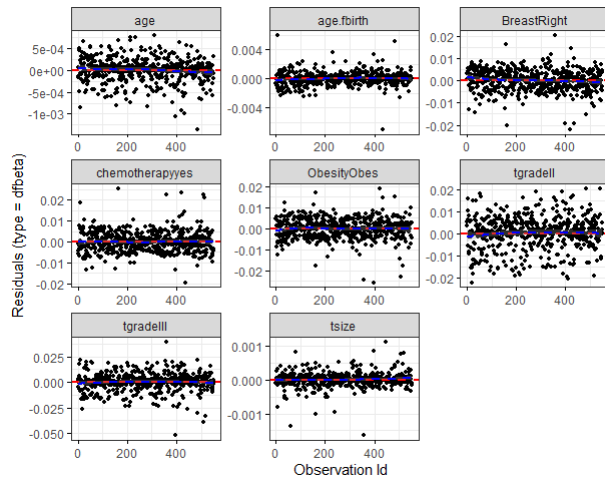


Figure 4. Variables' outlier for breast cancer data from the cox proportional model

As we can see in the plot for picture (5) on the next page, the age is the same for the three shapes. There is no difference between them. That means we can use age without log or root for the variable when we are dealing with cox proportional hazard model.

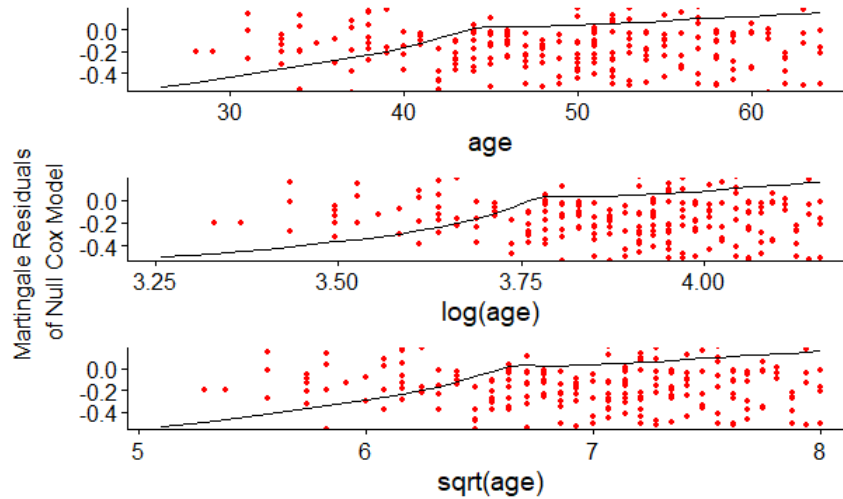


Figure 5. Using breast cancer data from the Cox proportional model to test for age linearity, which is a continuous variable

Also, from the plot below, the age at first birth is linear without log or root because there is no difference between the shapes. Tumor size also is linear without log or root since the shapes are the same. Finally, all continuous variables in this study are linear without log or root in the variable when we are dealing with cox proportional hazard model.

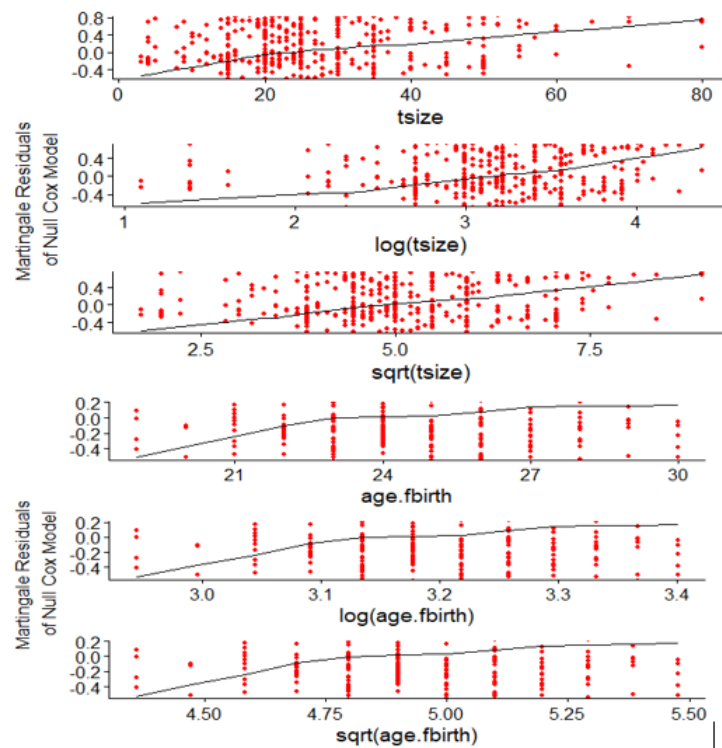


Figure 6. Using breast cancer data from the cox proportional model to test linearity for a continuous variable, age at first birth

That means all variables are time-independent. As a result, if the Cox Proportional Hazard assumptions were not violated, a Cox Proportional Hazard model could be used, as was done in this study.

4.2 Descriptive statistics

From the table(1) below, the average age of patients from the 551 cases had 49.06 years old. The value of skewness and kurtosis indicated that the data met the normality criteria. the tumor size mean was 26.28 mm. the average age of first birth was 24.38 days. Breast cancer patients have an average lifetime of 898.6 days or 2.5 years.

Table 1. summary of descriptive statistics for the breast cancer data with contentious variables

Summary	age	age.fbirth	tsize	pnodes
nobs	551	551	551	551
NAs	0	0	0	0
Minimum	26	19	3	1
Maximum	64	30	80	36
1. Quartile	43	23	18	2
3. Quartile	56	26	31.5	8
Mean	49.056261	24.381125	26.284936	5.987296
Median	50	24	24	4
Sum	27030	13434	14483	3299
SE Mean	0.3659	0.106789	0.580757	0.2484
LCL Mean	48.337528	24.171361	25.144164	5.499368
UCL Mean	49.774994	24.59089	27.425709	6.475224
Variance	73.769556	6.28357	185.840482	33.99802
Stdev	8.588921	2.506705	13.632332	5.830782
Skewness	-0.231555	0.058306	1.183121	1.635336
Kurtosis	-0.759385	-0.416676	2.004253	2.87058

Women patients with breast cancer registered an average of 58.8% deaths and 41.2% were censored. This suggests that breast cancer has a high death rate. Of the women patients, 36.3% were obese and 63.7% were not obese. Also women patients 61.52% were treated using medicine (chemotherapy) and 38.48% of them were treated in another way like surgery that means they did not use medicine as treatment. The tumor grades for breast cancer had an average of 34.85% for grade one (I), 39.02% for grade two (II) and 26.13% for grade three (III). stated from the table (7) below.

chemotherapy		Avarage
Yes	339.00	61.52
No	212.00	38.48

Breast		Avarage
Right	271.00	49.18
Left	280.00	50.82

Tumor Grade		Avarage
I	192.00	34.85
II	215.00	39.02
III	144.00	26.13

Obesity		Avarage
Obes	200.00	36.30
Not Obes	351.00	63.70

Censoring		Avarage
Died	324.00	58.80
A live	227.00	41.20

Figure 7. summary for the Category variables in the breast cancer data

As we can see from the table (2) for the next page, The average lifetime of Women taking medicine who had left breast cancer and they had not obese was 1000.6 days but the women who had obese were 1050.45 days. The average lifetime of Women who had not taken medicine and they had left breast cancer and they had not obese was 886.24 days but the women who had obese with left breast cancer were 837.82 days.

The average lifetime of Women taking medicine who had had right breast cancer and they had not obese was 907.11 days but the women who had obese were 838.26 days. The average lifetime of Women taking medicine who had had right breast cancer and they had not obese was 762.75 days but the women who had obese were 728.98 days.

Therefore the women who had taken medicine have an average lifetime longer than those who had not taken medicine. Also, the women who had right breast cancer have longer survival time according to the women who had left breast cancer.

Table 2. In the breast cancer data, the average time of women who are obese and have breast groups on the left or right

Average of time	Column Labels			
	Left		Left Total	
Row Labels	Not Obes	Obes		
no	886.2368421	837.8181818	875.3673469	
yes	1003.6	1051.447761	1021.214286	
Grand Total	956.9005236	998.6404494	970.1678571	
	Right		Right Total	Grand Total
Row Labels	Not Obes	Obes		
no	762.75	728.98	747.9385965	806.8443396
yes	907.1145833	838.2622951	880.3630573	955.9823009
Grand Total	849.36875	789.036036	824.6568266	898.600726

According to the table (3) on the next page, the average tumor size for women who had taken medicine and they had diagnosed with left breast cancer was 14.42 mm in grade one (I), 24.64 mm in grade two (II), and 41.07 mm in grade three (III), but women who had diagnosed with right breast cancer and they had taking medicine was 14.48 mm in grade one (I), 25.12 mm in grade two (II) and 41.62 mm in grade three (III).

From the table (3) on the next page, the average tumor size for women who had not taken medicine and had left breast cancer was 17.12 mm in grade one (I), 34.43 mm in grade two (II) and 41.59 mm in grade three (III), but women who had diagnosed with right breast cancer and who had not taken medicine was 18.03 mm in grade one (I), 26.69 mm in grade two (II) and 35.84 mm in grade three (III).

So that means a woman who has left breast cancer is better than a woman who has right breast cancer, also a woman diagnosed with breast cancer who is taking medicine is better than woman diagnosed with breast cancer who is not taking medicine or they use another way like surgery or hormone therapy.

Table 3. The average tumor size in breast cancer data for women with left or right breast cancer, whether they were receiving chemotherapy or not

Average of tsize	Column Labels				
	Left			Left Total	
Row Labels	I	II	III		
no	17.12195122	34.428	41.591	28.796	
yes	14.41791045	24.643	41.071	24.670	
Grand Total	15.44444444	27.815	41.25	26.114	
	Right			Right Total	Grand Total
Row Labels	I	II	III		
no	18.03125	26.974	35.837	27.807	28.264
yes	14.48076923	25.118	41.622	25.484	25.0472
Grand Total	15.83333333	25.794	38.513	26.4613	26.285

On the next page, the average age of the patients with left breast cancer who had not obese and they had taking medicine was 48.77 years old, but those who had not obese with right breast cancer and had taking medicine was 49.03 years old. While the average age of the patients with left breast cancer who had obese and they had taking medicine was 48.33 years old, but those who had obese with right breast cancer and they had taking medicine was 48.49 years old.

Also considering the table (4.4) on the next page, the average age of the patients with left breast cancer who had not obese and they had not taken medicine was 50.36 years old, but those who had not obese with right breast cancer and had taking medicine was 48.64 years old. While the average age of the patients with left breast cancer who had obese and they had taking medicine was 49.68 years old, but those who had obese with right breast cancer and they had taking medicine was 49.72 years old.

That means the women who are taking medicine were mostly younger than those who were not taking medication or they were using another treatment way like surgery.

Table 4. The average of age in breast cancer data for women with left or right breast cancer, whether they were receiving chemotherapy or not, plus obesity

Average of age	Column Labels			
	Not Obes		Not Obes Total	
Row Labels	Left	Right		
no	50.35526316	48.64063	49.57142857	
yes	48.76521739	49.03125	48.88625592	
Grand Total	49.39790576	48.875	49.15954416	
	Obes		Obes Total	Grand Total
Row Labels	Left	Right		
no	49.68181818	49.72	49.70833333	49.61792453
yes	48.32835821	48.4918	48.40625	48.70501475
Grand Total	48.66292135	49.04505	48.875	49.05626134

For the picture (8) on the next page, we found that according to chi-square there is not enough evidence to reject the null hypothesis from the relationship between chemotherapy and tumor grade, therefore there is no association between chemotherapy and tumor grade, because of the p-value (0.1217) is greater than 0.05. Also the table (8), we can't reject the null hypothesis from chemotherapy and breast groups, so there is no association between chemotherapy and breast groups since the p-value (0.1059) is greater than 0.05. That means they were independent of each other.

```
ss=xtabs(~sursom$chemotherapy+sursom$tgrade)
chisq.test(ss)
```

```
##
## Pearson's Chi-squared test
##
## data:  ss
## X-squared = 4.2126, df = 2, p-value = 0.1217
```

```
ss=xtabs(~sursom$chemotherapy+sursom$Breast)
chisq.test(ss)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ss
## X-squared = 2.6141, df = 1, p-value = 0.1059
```

Figure 8. Relationship between chemotherapy with tumor grade and breast

From the table (9) from R software, considering the chi-square test there is no relationship between chemotherapy and obesity because the p-value (0.4177) is greater than 0.05.

Also for the table (9) from R software, the p-value (0.7014) is greater than 0.05 so there is no association between breast groups and obesity.

Finally from the table (9) from R software, the p-value (0.843) is greater than 0.05 so there is no relationship between tumor grades and obesity. So that means all variables were independent of each other, therefore each variable was independent and identically distributed (IID).

```

ss=xtabs(~sursom$chemotherapy+sursom$Obesity)
chisq.test(ss)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  ss
## X-squared = 0.6569, df = 1, p-value = 0.4177

ss=xtabs(~sursom$Breast, sursom$Obesity)
chisq.test(ss)

##
## Chi-squared test for given probabilities
##
## data:  ss
## X-squared = 0.14701, df = 1, p-value = 0.7014

ss=xtabs(~sursom$Tgrade+sursom$Obesity)
chisq.test(ss)

##
## Pearson's Chi-squared test
##
## data:  ss
## X-squared = 0.34159, df = 2, p-value = 0.843

```

Figure 9. Relationship between chemotherapy and obesity and also between obesity and tumor grade

4.3 Survival Analysis

The dependent variable was described as the survival time and statistics of censoring. The independent variables were chemotherapy, Breast, Tumor Grade, Obesity, Age, Age at First Birth, and Tumor Size.

The variables represents like (age fbirth = Age at First Birth), (tsize = Tumor Size), (tgrade = Tumor Grade), (pnodes = Number of Positive Nodes) and the others are same words. The analysis software used was R studio and excel office.

The variables chemotherapy, Breast, Tumor Grade, Obesity, and Centering were factors for different levels. Regarding the picture (10) on the next page, we can see the structure in our breast cancer data.

```

sursom$chemotherapy=factor(sursom$chemotherapy)
sursom$tgrade=factor(sursom$tgrade)
sursom$Breast=factor(sursom$Breast)
sursom$Obesity=factor(sursom$Obesity)
sursom$Censoring=factor(sursom$Censoring)
str(sursom)

## 'data.frame':  551 obs. of  10 variables:
## $ chemotherapy: Factor w/ 2 levels "no","yes": 1 2 1 2 1 2 1 1 2 1 ...
## $ age         : int  53 64 38 34 39 36 41 57 48 53 ...
## $ age.fbirth  : int  23 20 19 24 24 25 28 19 26 19 ...
## $ tsize       : int  23 15 4 4 30 19 15 40 21 9 ...
## $ tgrade      : Factor w/ 3 levels "I","II","III": 2 1 1 1 2 1 2 2 2 1
## ...
## $ pnodes      : int  4 9 18 4 2 7 7 3 9 4 ...
## $ Breast      : Factor w/ 2 levels "Left","Right": 2 1 2 1 2 1 2 2 2 2
## ...
## $ time        : int  556 956 1489 1805 511 1803 1357 1535 509 1261 ...
## $ Obesity     : Factor w/ 2 levels "Not Obes","Obes": 2 1 1 2 2 1 2 2 1 1
## ...
## $ Censoring   : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 2 2 2 ...

```

Figure 10. Women breast cancer data structure from Mogadishu, Somalia

4.3.1 Kaplan Meier (KM)

Using the KM method defined in the chapter of methodology, the 551 cases for only women patients with breast cancer were studied.

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) \quad (28)$$

Where t denotes the current time, n denotes the number of patients at risk, d denotes the number of deaths at the current time, $\hat{S}(t)$ denotes cumulative survival probability estimator at time t .

The Kaplan Meier outcome was achieved by fitting the survival time employing median survival time as shown in the table (11) on the next page.

The median follow-up period for the breast cancer research was 1101 days or three years, also the upper and lower interval was 1030 days and 1165 days respectively according to this picture (11) on the next page.

```

> km=survfit(Surv(sursom$time, sursom$censoring)~1)
> km
Call: survfit(formula = Surv(sursom$time, sursom$censoring) ~ 1)

   n  events  median 0.95LCL 0.95UCL
551    324   1101   1030   1165
> summary(km)
Call: survfit(formula = Surv(sursom$time, sursom$censoring) ~ 1)

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
   92    551     1  0.9982 0.00181  0.9946  1.000
   93    550     1  0.9964 0.00256  0.9914  1.000
   96    548     1  0.9946 0.00314  0.9884  1.000
  110    545     1  0.9927 0.00362  0.9857  1.000
  114    544     1  0.9909 0.00405  0.9830  0.999
  135    543     1  0.9891 0.00443  0.9804  0.998
  149    539     1  0.9872 0.00479  0.9779  0.997
  152    538     1  0.9854 0.00512  0.9754  0.995
  163    537     1  0.9836 0.00543  0.9730  0.994
  168    535     1  0.9817 0.00572  0.9706  0.993
  189    531     1  0.9799 0.00600  0.9682  0.992
  192    530     1  0.9780 0.00627  0.9658  0.990
  197    529     1  0.9762 0.00653  0.9635  0.989
  202    528     1  0.9743 0.00677  0.9612  0.988
  203    527     1  0.9725 0.00701  0.9589  0.986
  222    524     1  0.9706 0.00723  0.9566  0.985
  249    518     1  0.9688 0.00746  0.9543  0.983
  253    515     1  0.9669 0.00768  0.9519  0.982
  262    514     2  0.9631 0.00810  0.9474  0.979
  284    508     1  0.9612 0.00830  0.9451  0.978
  302    506     1  0.9593 0.00850  0.9428  0.976
  305    504     1  0.9574 0.00869  0.9405  0.975
  311    502     1  0.9555 0.00888  0.9383  0.973
  314    501     1  0.9536 0.00906  0.9360  0.972
  ---

```

Figure 11. Kaplan Meier table

From the Kaplan Meier curve for the graph (12) below, the cumulative survival proportion in the women group who were taking medications or who were using chemotherapy as treatment tends to be significantly greater than in the women group who are using another treatment way like surgery, so that means women taking medicine will survive better than others. The survival proportion for women who are not obese is better or greater than women who are obese.

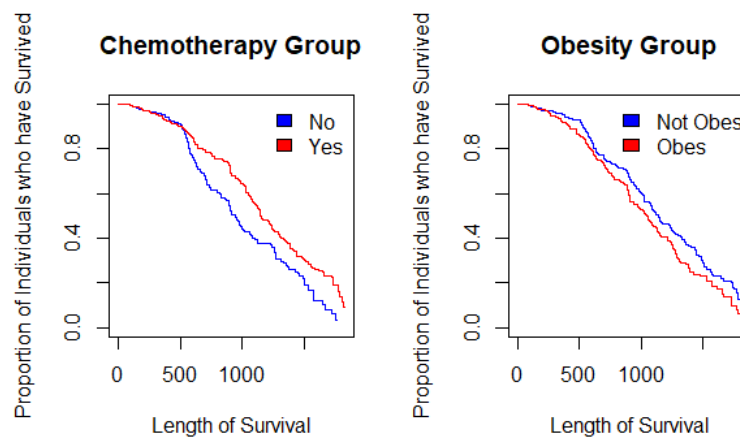


Figure 12. the Kaplan Meier curve for chemotherapy and Obesity groups

For the log-rank test result from the picture (13) on the next page, the survival time mean for women who were taking chemotherapy were 97.6 days and for the women who were not taking chemotherapy were 226.4 days. Considering the log-rank test there is enough evidence to reject the null hypothesis so we accept the alternative hypothesis that there is a difference between survival function according to those who were using chemotherapy and those who were not using chemotherapy because the p-value

is less than 0.05. In the log-rank test for obesity groups in women breast cancer patients, there is a significant difference between obese and non-obese women since the P-value is less than 0.05 for the table below.

```
> survdiff(Surv(sursum$time, sursum$censoring)~sursum$chemotherapy)
Call:
survdiff(formula = Surv(sursum$time, sursum$censoring) ~ sursum$chemotherapy)

              N Observed Expected (O-E)^2/E (O-E)^2/V
sursum$chemotherapy=no  212     127   97.6      8.82     13
sursum$chemotherapy=yes 339     197  226.4      3.81     13

Chisq= 13 on 1 degrees of freedom, p= 3e-04
> survdiff(Surv(sursum$time, sursum$censoring)~sursum$obesity)
Call:
survdiff(formula = Surv(sursum$time, sursum$censoring) ~ sursum$obesity)

              N Observed Expected (O-E)^2/E (O-E)^2/V
sursum$obesity=Not Obes 351     193   212      1.64     4.76
sursum$obesity=obes    200     131   112      3.10     4.76

Chisq= 4.8 on 1 degrees of freedom, p= 0.03
```

Figure 13. summary for log rank test for Chemotherapy and Obesity groups

As we can see the graph of Kaplan Meier curve (14) below, the survival proportion for women who had tumor grade one is better than who had tumor grade two or grade three, then the women who had tumor grade two somehow is better than in women had tumor grade three. Also, the survival proportion for women who had left breast cancer is better than those who had right breast cancer.

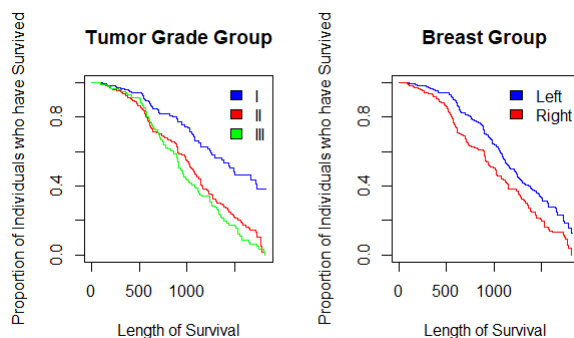


Figure 14. Plots of kaplan Meier for Tumor Grade and Breast

According to the log-rank test, there is a significant difference between tumor grades because the ($\chi^2 = 45.4$) with 2 degrees of freedom and the P-value is less than 0.05 for the table (15) below. Also, the log-rank test for breast groups whether right or left we reject the null hypothesis since ($\chi^2 = 16.8$) with one degree of freedom and the P-value is less than 0.05 for the picture (15) below.

4.3.2 Cox Proportional Hazard model

The Cox Proportional Hazard model described from the chapter of methodology was used to analyze the survival time for more than one variable.

$$Cox(x) = \frac{h(X)}{h_0(t)} = \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + x_p \beta_p) \quad (29)$$

```

> survdiff(Surv(sursom$time, sursom$Censoring)~sursom$tgrade)
Call:
survdiff(formula = Surv(sursom$time, sursom$Censoring) ~ sursom$tgrade)

          N Observed Expected (O-E)^2/E (O-E)^2/V
sursom$tgrade=I  192      61   117.1    26.86   42.76
sursom$tgrade=II 215     149   126.1     4.14    6.85
sursom$tgrade=III 144     114    80.8    13.65   18.30

Chisq= 45.4 on 2 degrees of freedom, p= 1e-10
> kmbr=survfit(Surv(sursom$time, sursom$Censoring)~sursom$Breast)
> survdiff(Surv(sursom$time, sursom$Censoring)~sursom$Breast)
Call:
survdiff(formula = Surv(sursom$time, sursom$Censoring) ~ sursom$Breast)

          N Observed Expected (O-E)^2/E (O-E)^2/V
sursom$Breast=Left 280     154    190     6.85   16.8
sursom$Breast=Right 271     170    134     9.73   16.8

Chisq= 16.8 on 1 degrees of freedom, p= 4e-05
> |

```

Figure 15. log rank test summary for tumor grade and breast groups

Model Selection

The likelihood ratio test was used to compare two models, a complete model, and a reduced model, using the ANOVA test. The null hypothesis is the reduced model has a good fit for the results. There are no parameters at all. ($\beta_1 = \beta_2 = \dots = \beta_p = 0$). The model is the same as an empty model with no explanatory variables, and it has a constant hazard in all classes.

According to the picture (16) below, the chi-square and p-value; if the p-value is significant, we will use the first model, which is the full model; if the p-value is not significant, there is no difference between the two models, the full model is not well-fitting, and we must build another model.

```

fullmodel=coxph(Surv(time, Censoring)~chemotherapy+Breast+Obesity+tgrade+age
+age.fbirth+tsize+pnodes, data = sursom)

reducedmodel|=coxph(Surv(time, Censoring)~1, data = sursom)

anova(cox8, cox0, test = "LRT")

```

Figure 16. model selection from breast cancer data from the cox proportional model

In the figure (17) on the next page, the p-value is very small, less than 0.05. The full model is useful for this research, and we can interpret the results from the full model, as well as use the final calculation in the research and predict the results.

The fitted cox proportional model is:

$$Cox(x) = \frac{h(X)}{h_0(t)} = \exp(\beta_1 \text{chemotherapy} + \beta_2 \text{Breast} + \beta_3 \text{Obesity} + \beta_4 \text{tgrade} + \beta_5 \text{age} + \beta_6 \text{age.fbirth} + \beta_7 \text{tsize} + \beta_8 \text{pnodes})$$

```

> fullmodel=coxph(Surv(time, Censoring)~chemotherapy+Breast+Obesity+tgrade+age
+
+age.fbirth+tsize+pnodes, data = sursom)
> reducedmodel=coxph(Surv(time, Censoring)~1, data = sursom)
> anova(fullmodel, reducedmodel, test = "LRT")
Analysis of Deviance Table
Cox model: response is Surv(time, Censoring)
Model 1: ~ chemotherapy + Breast + Obesity + tgrade + age + age.fbirth + tsize + pnode
s
Model 2: ~ 1
      loglik  Chisq Df P(>|chi|)
1 -1685.7
2 -1736.5 101.61  9 < 2.2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

Figure 17. Good fit model from breast cancer data from the cox proportional model

$$\begin{aligned}
 Cox(x) = \frac{h(X)}{h_0(t)} = & 0.7246 \text{chemotherapy yes} + 1.4040 \text{Breast Right} + \\
 & 1.2799 \text{Obesity(Obes)} + 1.7227 \text{tgradeII} + 1.7896 \text{tgradeIII} + 1.0162 \text{age} + \\
 & 1.0467 \text{age.fbirth} + 1.0178 \text{tsize} + 0.9762 \text{pnodes}
 \end{aligned}$$

Asymptotically, the likelihood ratio, Wald, and Score (Log rank) measures are the same, since the p-value is near to zero or equal according to the result (18) below.

```

> summary(cox8)
Call:
coxph(formula = surv(time, Censoring) ~ chemotherapy + Breast +
      obesity + tgrade + age + age.fbirth + tsize + pnodes, data = sursom)

n= 551, number of events= 324

      coef exp(coef) se(coef)      z Pr(>|z|)
chemotherapyyes -0.322080  0.724640  0.117726 -2.736 0.006222 **
BreastRight      0.339324  1.403998  0.115086  2.948 0.003194 **
Obesityobes      0.246820  1.279949  0.115287  2.141 0.032280 *
tgradeII         0.543875  1.722670  0.169713  3.205 0.001352 **
tgradeIII        0.581966  1.789553  0.217839  2.672 0.007550 **
age              0.016042  1.016171  0.006831  2.348 0.018856 *
age.fbirth       0.045611  1.046667  0.021859  2.087 0.036920 *
tsize            0.017634  1.017791  0.005275  3.343 0.000828 ***
pnodes          -0.024098  0.976190  0.010403 -2.316 0.020535 *
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
chemotherapyyes  0.7246  1.3800  0.5753  0.9127
BreastRight      1.4040  0.7123  1.1205  1.7593
Obesityobes     1.2799  0.7813  1.0211  1.6044
tgradeII        1.7227  0.5805  1.2352  2.4025
tgradeIII       1.7896  0.5588  1.1677  2.7426
age             1.0162  0.9841  1.0027  1.0299
age.fbirth      1.0467  0.9554  1.0028  1.0925
tsize           1.0178  0.9825  1.0073  1.0284
pnodes          0.9762  1.0244  0.9565  0.9963

Concordance= 0.649 (se = 0.017 )
Likelihood ratio test= 101.6 on 9 df,  p=<2e-16
Wald test               = 96.92 on 9 df,  p=<2e-16
Score (logrank) test = 102 on 9 df,  p=<2e-16

```

Figure 18. the results for the cox proportional model from breast cancer data

Table 5. A hazard ratio for the Cox Proportional Hazard Model result is shown in the table

Depend V(t,c):		all	HR (univariable)	HR (multivariable)
chemotherapy	no	212 (100.0)	-	-
	yes	339 (100.0)	0.66 (0.53-0.83, p<0.001)	0.72 (0.58-0.91, p=0.006)
age	Mean (SD)	49.1 (8.6)	1.03 (1.01-1.04, p<0.001)	1.02 (1.00-1.03, p=0.019)
age.fbirth	Mean (SD)	24.4 (2.5)	1.08 (1.04-1.13, p<0.001)	1.05 (1.00-1.09, p=0.037)
tsize	Mean (SD)	26.3 (13.6)	1.03 (1.02-1.03, p<0.001)	1.02 (1.01-1.03, p=0.001)
tgrade	I	192 (100.0)	-	-
	II	215 (100.0)	2.30 (1.70-3.11, p<0.001)	1.72 (1.24-2.40, p=0.001)
	III	144 (100.0)	2.75 (2.01-3.76, p<0.001)	1.79 (1.17-2.74, p=0.008)
pnodes	Mean (SD)	6.0 (5.8)	1.00 (0.99-1.02, p=0.580)	0.98 (0.96-1.00, p=0.021)
Breast	Left	280 (100.0)	-	-
	Right	271 (100.0)	1.58 (1.27-1.97, p<0.001)	1.40 (1.12-1.76, p=0.003)
Obesity	Not Obes	351 (100.0)	-	-
	Obes	200 (100.0)	1.28 (1.03-1.60, p=0.029)	1.28 (1.02-1.60, p=0.032)

Interpretation for Cox Model

Every Hazard Ratio represents a relative risk of death based on the comparison of one instance of a binary function to the other. According to the table for cox results (5) above, the following are interpretations of the Cox proportional hazard model that has eight independent variables dummy variable, so we interpret one by one.

- 1) The hazard ratio for the chemotherapy variable was 0.72, then the patients who received chemotherapy were (28%) less likely to die compared to patients who had not received chemotherapy but they were using another treatment, like surgery, remained for all other variables as constant.
- 2) A woman patient with right breast cancer has a hazard ratio of 1.40 when compared to a woman patient with left breast cancer as a reference category or as a baseline hazard. This indicates that a woman patient who has right breast cancer is (40%) more likely to die compared to a woman patient who has left breast cancer, adjusted for all other variables.
- 3) The Obesity variable with its non-obese as a reference category or as a baseline hazard has a hazard ratio value of 1.28. This indicates that a woman patient with breast cancer who is obese is (28%) more likely to die compared to a woman patient with breast cancer who is not obese, adjusted for all other variables.

-
- 4) Tumor grade II variable with tumor grade I category as reference or as a baseline hazard, hazard ratio value of 1.72. This means that the patient woman with breast cancer who had tumor grade II has a (72%) higher risk of dying compared to the patient woman with breast cancer who had tumor grade I, after adjusting for all other variables.
 - 5) Tumor grade III variable with tumor grade I category as reference or as a baseline hazard, hazard ratio value of 1.79. This means that the patient woman with breast cancer who had tumor grade III has a (79%) higher risk of dying compared to the patient woman with breast cancer who had tumor grade I.
 - 6) After adjusting for all other variables, the patient woman with breast cancer is (2%) more likely to die for every additional year of age, according to the hazard ratio for the age variable of 1.02.
 - 7) After adjusting for all other variables, according to the hazard ratio for age at the first birth variable of 1.05, the patient woman with breast cancer is (5%) more likely to die for every additional year of age at first birth.
 - 8) After controlling for all other variables, the patient woman with breast cancer is (2%) more likely to die for every additional millimeter in tumor size, according to the hazard ratio for tumor size variable of 1.02.
 - 9) The number of positive nodes in women with breast cancer has a negative regression coefficient with the value of -0.024098 and a hazard ratio value of 0.98. This indicates that a woman patient with breast cancer is (2%) less likely to die for every additional number of positive nodes a woman with breast cancer, after adjusting for all other variables.

5 Conclusion And Recommendations

5.1 Summary

Breast cancer can develop in any woman in Somalia, according to this research, since most patients were screened in the city of Mogadishu in Somalia. The key factors in speeding up breast cancer for females in Somalia were the age of the patient, age at first life birth, tumor size, tumor grade, breast site, and obesity. All of these risk factors are statistically significant. That means they are positively or negatively effective.

As a result, I discovered that all these eight variables were associated with the development of breast cancer, which matched the findings of this research. According to (Globocan 2020), the number of new cases of breast cancer in females is the highest of all cancers, the prevalence of breast cancer among females was 31.01%, and the mortality rate from breast cancer was 15% of females of all ages.

According to this research, the death rate in our sample was 59% that was 324 from the total of 551. That occurs because covid-19 causes a slew of issues, including a decrease in household income and a decrease in market labor, resulting in patients being unable to spend their earnings. As we can see, breast cancer is the most aggressive and lethal cancer in Somali women.

Since I only gathered data from women, the study had some limitations, such as inadequate data from hospital records, and the meager data obtained for females posed some questions. The patients had little knowledge of breast cancer and did not adhere to the medication very well, even though some of them did not follow the doctor's advice very well. Furthermore, the covid-19 had other limitations in the previous years until now, because most female patients came from poor people or rural areas, so they could only earn a living during the pandemic, and they couldn't afford to go to the hospital to pay for reception or medicine.

5.2 Conclusion

Breast cancer awareness programs in Mogadishu or other regions in Somalia can focus on early detection and provide more information to women. Cancer associations in Somalia must update their cancer registry and collect data on time to provide timely and accurate breast cancer prevalence and trends in the region. Even data from March 2020 until 2021 were supposed to be included in this report, but they have not yet reached into the cancer registry from the hospital.

The hospitals that can treat patients with breast cancer are located in the city of Mogadishu in Somalia, so the people are coming from very far away from the city. There is not enough health care in the rural areas and the towns in the regions in Somalia. As

a result, Somalia society should focus on the best ways to assist survivors in every way possible, including offering a helping hand, emotional support, sympathy, and hope to those women in their lives who have been diagnosed with breast cancer for good health and to intergrade with the community.

According to the tables and plots, Kaplan-Meier was used; the appropriate model in the breast cancer data was Cox's proportional hazard model because the independent variables did not depend over time. Log-rank was also used to compare different groups.

5.3 Recommendations

- 1) A greater focus over the breast cancer treatment passageway in Somalia, with an emphasis on developing opportunities for early screening at an early age.
- 2) To create a Somalia National Cancer Registry in Mogadishu and other large towns in Somalia's sub-counties to assist weak and poor people living in rural areas and the countryside.
- 3) More cancer-specific hospitals should be established in Mogadishu, as well as cancer-specific hospitals in Somalia's sub-counties.
- 4) Since there is currently no government policy on breast cancer screening, treatment, and monitoring, the study recommends that the Ministry of Health and Social Services draft up some policies.
- 5) All sub-counties in Somalia should be the focus of breast cancer awareness and prevention campaigns.

5.4 Further Research

- a) More additional predictors such as a family history of breast cancer, socioeconomic status, contraceptive use, and location should be added in the next research.
- b) Although financial issues, it's better to do research from all sub-counties in Somalia and collect the data from the urban, rural areas, and towns if it is possible.
- c) It is preferable to mobilize the entire country to raise awareness about cancer, particularly breast cancer.
- d) Within the Cox Proportional Hazard model, other mathematical models should be investigated by other investigator

Bibliography

- [Etikan et al., 2018] Etikan, I., Bukirova, K., & Yuvali, M. (2018). Choosing statistical tests for survival analysis. *Biom. Biostat. Int. J*, 7, 477-481.
- [Ferraz et al., 2017] Ferraz, R. D. O., & Moreira-Filho, D. D. C. (2017). Survival analysis of women with breast cancer: competing risk models. *Ciencia & saude coletiva*, 22, 3743-3754.
- [Jager, K, et al., 2008] Jager, K. J., Van Dijk, P. C., Zoccali, C., & Dekker, F. W. (2008). The analysis of survival data: the Kaplan–Meier method. *Kidney international*, 74(5), 560-565.
- [Goel, et al., 2010] Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, 1(4), 274.
- [Borgan, 2014] Borgan. (2014). Nelson–Aalen Estimator. *Wiley StatsRef: Statistics Reference Online*.
- [Colosimo, et al., 2002] Colosimo, E., Ferreira, F. V., Oliveira, M., & Sousa, C. (2002). Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4), 299-308.
- [Abadi, et al., 2014] Abadi, A., Yavari, P., Dehghani-Arani, M., Alavi-Majd, H., Ghasemi, E., Amanpour, F., & Bajdik, C. (2014). Cox models survival analysis based on breast cancer treatments. *Iranian journal of cancer prevention*, 7(3), 124.
- [Husain, et al., 2018] Husain, H., Thamrin, S. A., Tahir, S., Mukhlisin, A., & Apriani, M. M. (2018, March). The application of extended Cox proportional hazard method for estimating survival time of breast cancer. In *Journal of physics: Conference series* (Vol. 979, No. 1, p. 012087). IOP Publishing.
- [Kleinbaum, et al., 2010] Kleinbaum, David G and Klein, Mitchel, 2010. *Survival analysis*. Springer
- [Bewick, et al., 2004] Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *Critical care*, 8(5), 1-6.
- [Bewick, et al., 2020] J.A.M OTTIENO, 2020. *Survival Models*.
- [Kartsonaki, et al., 2016] Kartsonaki, Christiana, 2016. *Survival analysis*. *Diagnostic Histopathology*, 22, 7, pages 263–270. Elsevier

- [Rich, et al., 2010] Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3), 331-336.
- [Emmert-Streib, et al., 2019] Emmert-Streib, Frank and Dehmer, Matthias, 2019. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1, number 3, pages 1013–1038. Multidisciplinary Digital Publishing Institute
- [Carter, et al., 2009] Carter, R. E., & Huang, P. (2009). Cautionary note regarding the use of CIs obtained from Kaplan-Meier survival curves. *Journal of Clinical Oncology*, 27(2), 174-175.
- [Dakhil, et al., 2012] Dakhil, N. K., Mahdi, Y., & Al-A'bidy, M. A. M. (2012). Analysis of Breast Cancer Data using Kaplan–Meier Survival Analysis. *Journal of Kufa for Mathematics and Computer*, 1(6).
- [Etikan, et al., 2017] Etikan, İ., Abubakar, S., & Alkassim, R. (2017). The Kaplan-Meier estimate in survival analysis. *Biom Biostatistics Int J*, 5(2), 00128.
- [Cai, et al., 2014] Cai, Jianwen and Zeng, Donglin, 2014. *Cox Proportional Hazard Model*. Wiley StatsRef: Statistics Reference Online. Wiley Online Library
- [Tahtabasi, et al., 2014] Tahtabasi, M., Abdullahi, I. M., Kalayci, M., Ibrahim, I. G., & Er, S. (2020). Cancer Incidence and Distribution at a Tertiary Care Hospital in Somalia from 2017 to 2020: An Initial Report of 1306 Cases. *Cancer Management and Research*, 12, 8599.
- [Kantelhardt, et al., 2015] Kantelhardt, E. J., Cubasch, H., & Hanson, C. (2015). Taking on breast cancer in East Africa: global challenges in breast cancer. *Current Opinion in Obstetrics and Gynecology*, 27(1), 108-114.
- [Kantelhardt, et al., 2018] Adeloye, Davies and Sowunmi, Olaperi Y and Jacobs, Wura and David, Rotimi A and Adeosun, Adeyemi A and Amuta, Ann O and Misra, Sanjay and Gadanya, Muktar and Auta, Asa and Harhay, Michael O and others, 2018. Estimating the incidence of breast cancer in Africa: a systematic review and meta-analysis. *Journal of global health*. volume8, number1. International Society for Global Health.
- [Unger-Saldaña, et al., 2014] Unger-Saldaña, K. (2014). Challenges to the early diagnosis and treatment of breast cancer in developing countries. *World journal of clinical oncology*, 5(3), 465.
- [Tahtabasi, et al., 2020] Tahtabasi, M., Abdullahi, I. M., Kalayci, M., Ibrahim, I. G., & Er, S. (2020). Cancer Incidence and Distribution at a Tertiary Care Hospital in Somalia from 2017 to 2020: An Initial Report of 1306 Cases. *Cancer Management and Research*, 12, 8599.

-
- [Kartsonaki, et al., 2016] . Kartsonaki, Christiana, 2016. Survival analysis. Diagnostic Histopathology, volume 22, number 7, pages 263–270, publisher Elsevier.
- [Ferraz, et al., 2017] Ferraz, R. D. O., & Moreira-Filho, D. D. C. (2017). Survival analysis of women with breast cancer: competing risk models. *Ciencia & saude coletiva*, 22, 3743-3754.
- [Lukong, et al., 2017] Lukong, K. E., Ogunbolude, Y., & Kamdem, J. P. (2017). Breast cancer in Africa: prevalence, treatment options, herbal medicines, and socio-economic determinants. *Breast cancer research and treatment*, 166(2), 351-365.
- [Rezaianzadeh, et al., 2009] Rezaianzadeh, A., Peacock, J., Reidpath, D., Talei, A., Hosseini, S. V., & Mehrabani, D. (2009). Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran. *BMC cancer*, 9(1), 1-11.
- [Vanderpuye, et al., 2017] Vanderpuye, V., Grover, S., Hammad, N., Simonds, H., Olopade, F., & Stefan, D. C. (2017). An update on the management of breast cancer in Africa. *Infectious agents and cancer*, 12(1), 1-12.
- [Tahtabasi, et al., 2020] Tahtabasi, M., Abdullahi, I. M., Kalayci, M., Ibrahim, I. G., & Er, S. (2020). Cancer Incidence and Distribution at a Tertiary Care Hospital in Somalia from 2017 to 2020: An Initial Report of 1306 Cases. *Cancer Management and Research*, 12, 8599.
- [Hong, et al., 2017] Li, H. (2017). Survival Analysis for a Breast Cancer Data Set. *Advances in Breast Cancer Research*, 6(01), 1.
- [Bustan, et al., 2018] Bustan, M. N., Aidid, M. K., & Gobel, F. A. (2018, June). Cox Proportional Hazard Survival Analysis to Inpatient Breast Cancer Cases. In *Journal of Physics: Conference Series* (Vol. 1028, No. 1, p. 012230). IOP Publishing.
- [Tableman, et al., 2003] Tableman, M., & Kim, J. S. (2003). *Survival analysis using S* (Vol. 519, No. 001.6). Boca Raton: Chapman & Hall/CRC.
- [Tableman, et al., 2003] Lee, Elisa T and Wang, John, 2003. *Statistical methods for survival data analysis*. volume 476, publisher John Wiley & Sons