College of Natural Sciences

Department of Statistics

Survival Analysis of Recurrent Events on Women Breast Cancer: The case of Tikur Anbessa Specialized Hospital, Ethiopia

---

By: Tashome Fenta

A Thesis Submitted to Jimma Univeristy, College of Natural Sciences, Department of Statistics in Partial Fulfillment of the Requirements for the Degree of Master of Science in Biostatistics

February 28, 2020

Jimma, Ethiopia

Survival Analysis of Recurrent Events on Women Breast Cancer: The case of Tikur Anbessa Specialized Hospital, Ethiopia

By: Tashome Fenta

Advisor: Geremew Muleta (PhD Scholar)
Co-Advisor: Yasin Negash (M.Sc.)

February 28, 2020
Jimma, Ethiopia

Approval Sheet

This is to certify that the thesis titled "Survival Analysis of Recurrent Events on Women Breast Cancer: The case of Tikur Anbessa Specialized Hospital, Ethiopia" submitted in partial fulfillment of the requirement for the degree of Master of Science in Biostatistics to the college of natural science Jimma University, and is record of original research carried out by Tashome Fenta Biru, under my supervision and no part of the thesis has been submitted for another degree or diploma. The assistance and the help received during the course of this investigation have been duly acknowledged. Therefore, I recommended that would be accepted as fulfilling the thesis requirement.

Geremew Muleta (PhD Scholar)	.....	.....
Advisor	Signature	Date
Yasin Negash(M.Sc.)	.....	.....
Co-advisor	Signature	Date

As the members of the board of examiners of M.Sc. thesis open defense examination of Tashome Fenta Biru, we certify that we have read and evaluated the thesis and examined the candidate. We recommend that the thesis has been accepted as it fulfills the requirements for the degree of Master of Science in Biostatistics.

.....	.....	.....
Name of External Examiner	Signature	Date
.....	.....	.....
Name of Internal Examiner	Signature	Date
.....	.....	.....
Name of Chairman	Signature	Date

STATEMENT OF THE AUTHOR

As author of this research study, I declare that the thesis is a result of my genuine work, support of my supervisors and help hands of other individuals. Thus, all those had who participated in the study and sources of materials used for writing this thesis have been duly acknowledged. I have submitted this thesis to Jimma University as a partial fulfillment for the requirements of Degree of Master of Science in Biostatistics. The library directorate of Jimma University can deposit the copy of the thesis in the university library so that students and researchers can refer it. Moreover, I declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate and/or to get prove of societys problems.

Any brief quotations from this thesis are allowed without requiring special permission if an accurate acknowledgement and citation (after publication) of the source is made. In all other instances, however, permission must be obtained from the author.

Tashome Fenta ....

Name Signature Date

Acknowledgment

First and for most, I would like to extend my unshared thanks to the almighty God for providing me the opportunity for what I have achieved and for his mercy. I would like to express sincere appreciation to my Thesis advisor, Geremew Muleta (PhD Scholar), for sharing his substantial experience to do this research in the expect way and giving wonderful personality through the time without any reservation time. I would like to express my heartfelt thanks to my co-advisor, Yasin Negash (M.Sc.), for his valuable comments and suggestion.

My sincere thanks are due to the Tikur Anbessa Specialized Hospital for giving me the data for study. My special appreciation and acknowledgment go to Assosa University for offering me the opportunity and financial support during my study at Jimma University.

Finally, I take this opportunity to sincerely express my gratitude to my beloved family who are the source of pride and encouragement throughout my life. I am thankful to my father, Fenta Biru , my mother, Ayantu Nemera, my wife Selamawit Endale and to all my Brothers, my sisters and all my best friends for your unconditional love and supports.

Contents

List of Tables

List of Figures

Acronyms

ACS..... American Cancer Society

AJCC.....American Joint Committee on Cancer

Cox-PH.....Cox Proportional Hazard model

FMOH.....Federal Ministry of Healthy

GLOBOCAN.....Global Burden of Cancer

LCV.....Likelihood cross-validation criterion

IARC.....International Agency for Research on cancer

SE.....Standard Error

SSA.....Sub-Saharan Africa

SNNP.....Southern Nation Nationality People of Ethiopia

NCI.....National Cancer Institute

TASH.....Tikur Anbessa Specialized Hospital

WHO.....World Health Organization

Abstract

Background: Breast cancer is the most commonly diagnosed cancers worldwide. It is a cancer that develops from breast tissue and most common invasive cancer in women. Recurrent events data have been increasingly important in clinical studies where individuals experience an event more than once and it is a major clinical indicator, which represents the principal cause of breast cancer-related deaths.

Objective: The aim of the study was to investigate determinants of the recurrence of breast cancer.

Methodology: To reach the aim, 421 women with breast cancer were included in the study based on data taken from medical record card of patients enrolled starting from 1^{st} January 2013 to 30^{th} January 2019. A retrospective study has been applied to obtain data on women breast cancer that recorded in oncology department of Tikur Anbessa Specialized Hospital. Unmeasured shared similarities due to the impact of multiple events were modeled using a random effect. Cox-PH model and Shared frailty model were used to identifies which factor was significantly affect the recurrence of breast cancer.

Results: From the total of 997 recurrent events, about 609 (61.1%) of them experienced recurrence of breast cancer. The shared log-normal frailty model was chosen as the best fit for this breast cancer data set based on the value of Likelihood cross-validation criterion. From the result of shared log-normal frailty model age, stage of breast cancer, tumor size, histology grade, breast feeding and oral contraceptives were significantly associated with recurrence of women breast cancer.

Conclusion and recommendation: The result of shared log-normal frailty model shows that the stage (II, III, IV), tumor size ((3-5) cm, >5 cm), histology grade (poorly differentiated) and oral contraceptive were significantly increases the risk of recurrence of breast cancer. While, breast feeding was significantly decreases the risk of recurrence of breast cancer. It is recommended that policy maker, ministry of health and Tikur Anbessa Specialized Hospital are expected to make interventions based on these hazardous groups for recurrence of breast cancer.

Key words: ***Breast cancer, Counting Approach, Recurrent events, Shared frailty model***

1 Introduction

1.1 Background of the Study

Breast cancer is a cancer that develops from breast tissue and most common invasive cancer in women (Adesina *et al.*, 2013). It starts when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump. Breast cancer occurs almost entirely in women, but men can get breast cancer (Adams *et al.*, 2017).

Breast cancer is the most-frequently diagnosed cancer and the leading cause of cancer death among women worldwide, with an estimated 1.7 million cases and 521,900 deaths (IARC, 2013). It is a major life threatening and has become the major public health problem of great concern and the most common cause of cancer death among women in less developed countries (WHO, 2015). This cancer accounts for 25% of all cancer cases and 15% of all cancer deaths among women. Both in developed and developing countries breast cancer is a major health problem account for about one-half of all cancer cases and more than 324, 300 deaths occurred respectively (FMOH, 2015). It becomes an issue of public health in both developed and developing nations because of its high incidence-prevalence, the over-burdened health system and direct medical expenditure. Thus, breast cancer is the most common leading cause of cancer death problems in Africa (Parkin *et al.*, 2012).

In Ethiopia breast cancer is the most leading cancer occurring among women. It is estimated that around 9,900 Ethiopian women have breast cancer with thousands of more cases unreported as women living in rural areas often seek treatment from traditional healers before seeking help from the government health system. A retrospective study conducted in TASH from 1997-2012 indicates that of total 16,622 new cancer cases registered 3460 were new cases of breast cancer with prevalence of 20.8% and approximately 216 cases per annual (Abate *et al.*, 2016). The incidence of the breast cancer increases from year to year. Factors such as age, lymph node status, stage, histology grade, and hormone therapy have statistical significant effect on the recurrence of women breast cancer (Ahmedin *et al.*, 2012).

Repeated event processes, where individual subjects or units under consideration experience the same or different types of events more than once over time are called recurrent events. In many scientific investigations, the outcome variable of interest is a recurrent event. Recurrent event data are ubiquitous across a great range of diverse fields such as medicine, public health, insurance, social science, economics, manufacturing and reliability (Cook and Lawless, 2007). A logical objective for such kind of data is to assess the relationship of relevant predictors to the rate in which events are occurring, allowing for multiple events per subject (David *et al.*, 2005). Recurrent events are also observed in breast cancer patients. The recurrence of breast cancer is a major clinical indicator, which represents the principal cause of breast cancer-related deaths. It is occurred when cancer cells are detected following the initial treatment with surgery, radiotherapy or chemotherapy. Treatment options for recurrence of breast cancer vary depending on the previous treatment, the location of the recurrence, and the overall condition of the patient (Pan *et al.*, 2005).

Recurrence of the disease may depend on the extent of the disease, primary treatment and performance status/ co morbidity of the patient. The most simple analysis approach in a recurrent event setting is to count the events observed within a given time period. These counts may follow a Poisson, a quasi-Poisson or a negative binomial distribution (Wang *et al.*, 2009). Whenever patients are not all fully observed but are subject to an underlying censoring mechanism, analysis strategies for event times should be preferred over simple counting approaches. This situation is much more common in clinical application, but our focus lies on models for event times rather than on counting models.

In this study, the researcher applied survival analysis since it addresses the limitation of classical regressions like logistic and linear regressions. The Cox proportional hazard model is one of the common approaches to the analysis of time to event data. That is, conditional on the covariates, every individual has the same risk of experiencing an event such as disease recurrence. But the common Cox model only considers the time until the first occurring event meaning that all events after the first are neglected. The most frequently applied analysis method for recurrent time-to-event data is the model by Andersen and Gill (1982) which is based on the common Cox proportional

hazards model (Cox, 1972). The Andersen-Gill model assumes independence between all observed event times irrespective whether these event times correspond to the same patient or to different patients. In addition, The frailty is included in the model to account for variability due to unobserved subject-specific factors that are otherwise unaccounted for by the other predictors in the model. These unobserved subject-specific factors can be a source of within-subject correlation. We use the term shared frailty to indicate that observations are clustered by subject and each cluster (i.e., subject) shares the same level of frailty.

The shared frailty model was extended by Pickels *et al* (1994) and Yashin *et al* (1995) to allow different but correlated frailties among observations within a group. Frailty models account for unobserved heterogeneity that occurs because some observations are more prone to failure, and therefore more frail than others in a data set (Pickels *et al.*, 1994; Yashin *et al.*, 1995). Therefore, this study aims to review survival analysis techniques with frailty models and apply these methods on breast cancer data set to investigate determinants of recurrence of breast cancer and also to examine whether unobserved heterogeneity or unobserved covariates help to explain the recurrence of breast cancer using shared frailty model.

1.2 Statement of the Problem

Breast cancer is the most commonly diagnosed cancer in women both in the developed and less developed world (WHO, 2015). It has the highest incidence rate of all cancers in women worldwide around 1.67 million new cases and cause of over 500,000 deaths annually (Ferlay *et al.*, 2015). Breast cancer is the most common leading cause of cancer death problems in Africa including Ethiopia and Sub-Saharan countries (Tefera *et al.*, 2016; Jemal *et al.*, 2011). It is the most common cancer death occurs among women in developing countries, particularly in SSA, and its survival tends to be poor in this region because of a combination of a late stage at diagnosis and limited access to timely and standard treatment (Parkin *et al.*, 2012). A principal factor in decreasing survival rate in breast cancer patients is recurrence. Recurrence risk is measured by a collection of breast cancer natural history, prognostic, anatomic, biological factors, and the type of treatment (Cardoso *et al.*, 2012). Despite, the government concern on

a cancer issue in order to reduce the recurrence, incidence, mortality and the survival of women, breast cancer were still emergence in Ethiopia.

This study addresses research problem using the shared frailty models. Therefore, this study provides an extension of Cox PH model called the frailty model, taking into account any extra heterogeneity present in the data. In a different context, correlated data may come from recurrent events, i.e., events that occur several times within the same subject during the period of observation. Ignoring the existence of heterogeneity will produce biased parameter estimates and inconsistent standard errors in survival analysis.

Few authors have tried to determine the factors associated with recurrence of breast cancer using Cox proportional hazard model but no one has taken into consideration of unobserved heterogeneity in the data (Cheng *et al.*, 2012). Therefore, the researcher have employed a shared frailty model to investigate the factors associated with recurrence of breast cancer taking into account the heterogeneity. Generally, since we did not yet found study conducted on recurrence of women breast cancer with counting process at national level and the cases under study is found to be really a predominant issue, it happened to be a reason to conduct this study.

Thus, the study has attempted to answer the following basic research questions:

- Which factors significantly affect the recurrence of women breast cancer?
- Which model is best model for analyzing the predictors of recurrence of women breast cancer?

1.3 Objectives of the Study

1.3.1 General Objective

The aim of the study was to investigate determinants of recurrence of women breast cancer The case of Tikur Anbessa Specialized Hospital, Ethiopia.

1.3.2 Specific Objectives

The specific objectives of the study are:

- To explore appropriate model for recurrence of breast cancer among Cox PH, shared gamma frailty and shared log-normal frailty models.
- To identify the significant risk factors associated with the recurrence of breast cancer.

1.4 Significance of the Study

Studying the recurrence of women breast cancer is a mechanism of overcoming the problem of healthy in the society by identifying factors associated with recurrence of the disease. Modeling the recurrent events has been used to assess the possible risk factors for the recurrence of women breast cancer and on the basis of this model different prevention as well as treatment programs has been provided for the women. The result of this study would help to reduce the recurrence of breast cancer by giving awareness for the society on the factors that increase the probability of recurrence of the breast cancer and also it helps policy makers in making policy specially policy that are related to health.

2 Literature Review

2.1 Overview of Breast Cancer Recurrence

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow into surrounding tissues or spread (metastasize) to distant areas of the body. The disease occurs almost entirely in women, but men can get it, too. The normal female breast is made up mainly of lobules (milk producing glands), ducts (tiny tubes that carry the milk from the lobules to the nipple), and stroma (fatty tissue and connective tissue surrounding the ducts and lobules, blood vessels, and lymphatic vessels). Most breast cancers begin in the cells that line the ducts (ductal cancers). Some begin in the cells that line the lobules (lobular cancers), while a small number start in other tissues (ACS, 2016).

Breast cancer is the most frequently diagnosed and leading cause of cancer related deaths among females worldwide. According to estimates from GLOBOCCAN 2012, there were 100,000 cases and 49,000 deaths due to female breast cancer in the African region (Ferlay *et al.*, 2015). International records suggest that about 30% of women will develop recurrence after treatment for primary breast cancer, figures for early stage disease being lower (Geurts *et al.*, 2017). Despite this increasing cancer burden in Africa, it continues to receive a low public health priority, mainly due to limited resources and other pressing public health problems, including communicable diseases such as Human Immunodeficiency Virus (HIV) /(Acquired Immunodeficiency Syndrome (AIDS) infection, malaria, and tuberculosis (Ahmedin *et al.*, 2012).

In sub-Saharan Africa, breast cancer is responsible for one in four diagnosed cancers and one in five cancer deaths in women. Although its emerging public health importance, incidence rates are still generally low in Africa, presumably below 35 per 100,000 women in most countries as compared to over 90 - 120 per 100,000 women in most European or North American countries. Precise incidence figures in Africa, however, are lacking given the absence of cancer registration in most countries. Recent data estimate that in 2012, 94,000 women developed breast cancer and 48,000 died from it in sub-Saharan Africa. It has been estimated that by 2050, the incidence of

Breast cancer in Africa will be double of current 2012 estimate (Brinton *et al.*, 2014).

A study done on global burden of cancer showed 2.4 million women were diagnosed With 523,000 related deaths due to breast cancer in 2015 (GLOBCAN, 2017). Approximately 60% of deaths due to breast cancer occur in developing countries (da Costa Vieira *et al.*, 2017).

2.2 Risk factors of recurrence of Breast Cancer

There are many known cancer causes including lifestyle factors, such as tobacco use and excess body weight, and non-modifiable factors, such as inherited genetic mutations, hormones, and immune conditions. At least 42% of newly diagnosed cancers in the United State about 729,000 cases in 2018 are potentially avoidable, including 19% that are caused by smoking and 18% that are caused by a combination of excess body weight, physical inactivity, excess alcohol consumption, and poor nutrition. Like most cancers, many of the known breast cancer recurrence risk factors includes family history of breast cancer; personal history of breast cancer; early menarche ($\leq$12 years); late menopause ($\geq$ 55 years); aging; alcohol; late age at first full-term pregnancy ($\geq$ 30 years); never breastfed a child; recent oral contraceptive use; high fat diet; tobacco smoke; obesity (postmenopausal); recent and long-term use of hormone replacement therapy; high-dose radiation to chest; lack of physical activity (ACS, 2018).

Age:- The risk of developing breast cancer increases as once get older. About 1 out of 8 invasive breast cancers are found in young women less than 45, while about 2 of 3 invasive breast cancers are found in women age 55 or older. Also younger women, particularly those under the age of 35, have a higher risk of recurrence of breast cancer. Breast cancer incidence and death rates generally increase with age (Dignam *et al.*, 2009). During 2010-2014, the median age at the time of breast cancer diagnosis was 62. This means that half of women who developed breast cancer were 62 years of age or younger at the time of diagnosis (Howlader *et al.*, 2017).

Alcohol Consumption:- Alcohol consumption increases the relative risk of breast cancer in women for about 7% to 10% (Chen *et al.*, 2011). While only a few studies have been done on drinking alcohol and the risk of recurrence, drinking even a few alcoholic beverages per week (three to four drinks) increased the risk of breast cancer

coming back in women who'd been diagnosed with early-stage disease. Women who have 2-3 alcoholic drinks per day have a 20% higher risk of breast cancer recurrence compared to non-drinkers (Liu *et al.*, 2015).

Smoking:- Smoking may increase the risk of breast cancer recurrence, particularly long term, heavy smoking and among women who start smoking before their first pregnancy. The 2014 US Surgeon General's report on smoking concluded that there is "suggestive but not sufficient evidence that smoking increases the risk of breast cancer recurrence (ACS, 2015). Women who initiated smoking before the birth of their first child had a 21% higher risk of breast cancer recurrence than non smoker women (Gaud *et al.*, 2013).

Obesity:- Obesity increases the risk of postmenopausal breast cancer recurrence as stated on American Institute for Cancer Research. The risk is about 1.5 times higher in overweight women and about 2 times higher in obese women than in lean women. It is a major risk factor for breast cancer recurrence and morbidity in both pre-menopausal and postmenopausal women (Vecchia *et al.*, 2011).

Breastfeeding:- Most studies suggest that breastfeeding for a year or more slightly reduces a women overall risk of breast cancer recurrence, with longer duration associated with greater risk reduction (Faupel-Badger *et al.*, 2013). In a review of 47 studies in 30 countries, the risk of breast cancer recurrence was reduced by 4% for every 12 months of breastfeeding(Britt *et al.*, 2007). Women treated for breast cancer that previously breast fed their babies have lower risk of recurrence than those who did not (Anderson *et al.*, 2014).

Oral Contraceptives:- Oral contraceptives were associated with recurrence of breast cancer (Lu Y, *et al.*, 2011; Saxe *et al.*, 1999). This implies that oral contraceptives pose a higher risk of breast cancer recurrence.

Treatments Taken:- Breast cancer typically is detected either during a screening examination, before symptoms have developed, or after symptoms have developed, when a woman feels a lump. Taking into account tumor characteristics, including size and extent of spread, as well as patient preference, treatment usually given to women breast cancer are surgery, radiotherapy, hormone therapy and chemotherapy. The recurrence of women breast cancer affected by treatment they receive (ACS, 2017). There are

different treatment options for breast cancer such as surgery which is recommended for most women with early stage combined with other treatments to reduce the risk of recurrence. It includes radiation therapy, chemotherapy, hormonal therapy and targeted therapy. Systemic therapies like chemotherapy, targeted therapy and hormonal therapy are primary treatment options for patients with metastatic diseases. The treatment options are decided by both patients and physician based on the clinical stage and biological characteristics of the cancer, the age of the patient and considering the risks and benefits associated with each option (ACS, 2015).

Stages:- Stage of breast cancer has been categorized into four stages (I, II, III and IV), using the tumor, node, metastasis system (AJCC, 2016). The recurrence of women breast cancer are significantly affected by the stage at diagnosis of breast cancer. Stage at diagnosis has a significant effect on the recurrence of breast cancer(Demicheli *et al.*, 2010). Patients with advanced-stage disease had much more hazard rates than those with other stage disease (Dignam *et al.*, 2009). From the results of these two studies the hazard rate for recurrent events of women with breast cancer was greatest as the stage increases.

Tumor Size:- Tumor size has been categorized as $\leq$ 2 cm, 3-5 cm and >5cm (AJCC, 2016). Tumor size has a significant effect on the increasing the risk of breast cancer recurrences (Mauguen *et al.*, 2013; Rondeau *et al.*, 2007). In general, the larger the tumor, the greater the chance of recurrence.

Histology Grade:- Histology grade (I, II and III) at diagnosis was significantly affected the recurrence of women breast cancer (Baghestani *et al.*, 2015). Histology grade at diagnosis was significantly affected the recurrence of women with breast cancer and the hazard rate was high for women in poorly differentiated grade(grade III) as compared to women in well differentiated grade(grade I)(Dignam *et al.*, 2009; Mauguen *et al.*, 2013).

2.3 Overview of Recurrent Models

Therneau (2000) and Clayton (1978) discusses recurrent event and compares them to generalized linear models, such as Poisson and logistic regression. A number of proportional hazard type models have been proposed for use with recurrent event data

(Therneau, 2000) and (Clayton, 1978). The counting process approach is used when recurrent events are treated as identical. If all recurrent events on the same subject are treated as identical, then the analysis required of such data is different than what is required if either recurrent events involve different disease categories and/or the order that events reoccur is considered important (Anderson *et al.*, 1993).

In Cox and parametric models, the hazard function may depend on unknown or no measurable factors which can cause the regression coefficients estimated from such models to be biased. In order to overcome the problem and better model recurrence of diseases, the frailty models were introduced. In fact, these models are used to explain the random variation of survival function due to unknown risk factors, such as genetic factors and numerous environmental factors. Semi-parametric inference for frailty models was introduced by (Klein *et al.*, 1992; Nielsen *et al.*, 1992). As suggested by Gill (1982) they use an EM algorithm applied to the Cox partial likelihood (Gill, 1982). Hastie and Tibshirani (1993) proposed a general model with time varying coefficients and suggested estimation through penalized partial likelihood. Therneau and Grambsch (2000) noted a link between the gamma frailty model and a penalized partial likelihood. In this study, we penalize the hazard function while Therneau and Grambsch (2000) penalize the frailties.

Ullah (2014) studied on the statistical modeling of recurrent events for sport injuries and compared the survival models given below: Andersen-Gill (A-G), Frailty, Wei-LinWeissfeld total time (WLW-TT) marginal, Prentice-Williams Peterson gap time (PWP-GT) conditional models for the analysis of recurrent injury data (Ullah *et al.*, 2014). The extensions of frailty model like shared frailty model: cure frailty model for a mixture of susceptible and insusceptible subjects for the event of interest, Nested frailty model when the data are clustered at several hierarchical levels, and Joint frailty model for the joint analysis of recurrent events and death are used for modeling recurrent events and death of cancer patients. They perform a semi-parametric penalized likelihood approach for parameter estimation with the different extensions of the simple shared frailty models to analyze recurrent events not a comparison (Rondeau *et al.*, 2010).

3 Data and Methodology

3.1 Description of the study area

The data set employed under this study is recurrence of breast cancer data which were collected from TASH. It is the biggest referral public hospital in Ethiopia and training center of health professionals. The hospital gives service to population of Addis Ababa city and its surroundings majorly, but patients come to the hospital from all over Ethiopia. The hospital gives service as both inpatient and ambulatory follow up. It starts an organized oncology service in 1998 Ethiopian calendar. The cancer unit at the TASH provides chemotherapy, radiation therapy, hormone therapy and other supportive and palliative cares. It is the main center for cancer registry, early detection, prevention, standard treatment and palliative care in Addis Ababa.

3.1.1 Study Population

This is a retrospective study aims to determine the recurrence of breast cancer based on hospital registry in TASH, oncology center. The population of this study was all women with breast cancer who had been registered at TASH starting from 1^{st} January 2013 to 30^{th} January 2019. All the data had been carefully reviewed from the registration log book and patients' registration card; if any inadequate information counters it has been checked from the file and excluded from analysis if proven to be inadequate.

3.1.2 Inclusion and Exclusion Criteria

Inclusion Criteria:- All breast cancer women patients came for the reason of cancer recurrence and registered with full information including in the registration log book or in the patients identification card were considered to be eligible for the study.

Exclusion criteria:- Patients with insufficient information about one of the vital variables either in the registration book or in the card were not eligible. Also the patients lost from the study without starting any breast cancer recurrence treatment was not included.

3.1.3 Data Collection Procedure

Ethical permission has been obtained from the oncology department of the TASH. The hospital based data collected by trained enumerator and principal investigator. This study incorporates secondary data.

3.1.4 Data Structure for Modeling Recurrent Events on Breast Cancer

Breast Cancer data has been checked carefully to identify first, second, third and additional times of time to recurrence on breast cancer (i.e. recurrent event data).

Table below illustrates data structures required for modeling recurrent events of time to recurrence of breast cancer. The time from the start of the follow-up (1^{st} January 2013) to first time to recurrence, second time to recurrence, third time to recurrence and the time to the last censorship in patients who did not have recurrence time of breast cancer was considered to model recurrent event data. On the other hand, data from breast cancer patients with both one and more time to recurrence were accounted when modeling recurrent event data. The time interval for recurrent events of time to recurrence for each patient was given by the difference between two successive recurrence times of breast cancer patient.

Table 3.1: Data Structure for recurrence of Breast Cancer

ID	Start	Stop	time	Event status	Tumor size	Treatment taken
1	0	8	8	1	1	Combination of ≥2
1	10	16	6	1	2	Radiation
1	23	32	9	0	1	Chemotherapy
2	0	9	9	1	3	Combination of ≥2
2	13	19	6	1	3	Surgery
2	21	31	10	0	2	Hormonal
3	0	7	7	1	1	Surgery
3	14	24	10	0	3	Chemotherapy

A pair of variable (start, stop) is used to define the time interval of the breast cancer recurrence.

Start:- the start time of the interval (in months). The entry of the recurrent events data would be considered from the day that the women starts diagnosis after breast becomes abnormal. The start time is generally equal to zero for the 1^{st} recurrence of breast cancer.

Stop:- the time (in months) at which the event occurs or the time of censoring (death, drop-out, or the end of study).

Time:-the number of months at risk that is calculated from the columns 'Start' and 'Stop'. This structure of the data can used to fit Andersen and Gill model and frailty model.

An indicator variable, Event, represents the status of every observation "1" if recurrence observed and "0" if the event is censored due to several reasons. For simplicity it is assumed that the censoring mechanism was independent of the recurrent event process in this study.

Among the breast cancer patients there are women that are dropped treatment, died, does not re-occur at the end of study. This means that the type of the survival data is random right censored.

3.2 Variables in the Study

3.2.1 Response Variable

The response variable for this study was time to recurrence of women breast cancer which measured in months.

3.2.2 Explanatory Variables

The predictor variables which assumed to influence the recurrence of women breast cancer were given in below:-

1. Age ($\leq$ 35, 36-49, $\geq$ 50)
2. Alcohol(No, Yes)

3. Smoking Status (No, Yes)
4. Residence (Urban, Rural)
5. Histology grade (I, II, III)
6. Breastfeeding (No, Yes)
7. Treatment taken (Chemotherapy, Radiation, Surgery, Hormone Therapy, Combination of the two or more)
8. Stage (I, II, III, IV)
9. Oral contraceptives (Not used, used)
10. Obesity (underweight, normal, overweight)
11. Tumor size ($\leq$ 2 cm, 3-5 cm, $>$ 5cm)
12. Family History (No, With)
13. Menopause status (pre, post)
14. Marital status (Single, Married, Widowed, Divorced)

3.2.3 Description of variables

- **Breast cancer recurrence:-** Recurrent breast cancer is breast cancer that comes back after initial treatment. It is not a new cancer; it is the same cancer the person originally had.
- Family History of breast cancer: having a family who had breast cancer or not.
- **Treatments Taken:-** Taking into account tumor characteristics, including size and extent of spread, as well as patient preference, treatment usually given to women with breast cancer are surgery, radiotherapy, hormone therapy and chemotherapy.

- **Hormone therapy:-** Women with hormone receptor positive breast cancer may reduce their risk of recurrent breast cancer by taking hormone therapy after their initial treatment. Hormone therapy may continue for at least five years.

- **Chemotherapy:-** For women with breast cancer who have an increased risk of cancer recurrence, chemotherapy has been shown to decrease the chance that cancer will recur, and those who receive chemotherapy live longer.

- **Radiation therapy:-**Women who have had a breast sparing operation to treat their breast cancer and those who had a large tumor or inflammatory breast cancer have a lower chance of the cancer recurring if they are treated with radiation therapy.

3.3 Methods of Data Analysis

3.3.1 Descriptive Statistics

The description of survival data utilizes non-parametric methods to compare the survival functions of two or more groups and Kaplan-Meier plot(s) would be employed for this purpose (Kaplan and Meier, 1958). The frequency distribution table was also used to summarize the data based on the study variables.

3.3.2 Survival Data Analysis

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. By time, mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs. By event, we mean death, disease incidence, relapse from remission, recovery or any designated experience of interest that may happen to an individual (Aalen *et al.*, 1980).

The use of survival analysis, as opposed to the use of other statistical method, is most important when some subjects are lost to follow up or when the period of observation is finite certain patients may not experience the event of interest over the study period. Most survival analyses consider a key analytical problem of censoring.

In essence, censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly. The censoring used in this paper is right censoring. Survival time is said to be right censored when it is recorded from its beginning to a defined time before its end time.

In reality right censoring can occur due to the following reasons:

- Death from unrelated causes
- Loss of follow-up
- Termination of study

Survival Function:- The survivor function is the probability that the survival time of a randomly selected subject is greater than or equal to some specified time. Thus, it gives the probability that an individual surviving beyond a specified time. The distribution of survival time is characterized by survivor ship, probability density function and hazard function.

Let T be a random variable associated with the survival times and t be the specified value of the random variable T and f(t) be the underlying probability density function of the survival time T. The survivor function, S(t), is given by:-

$$S(t) = P(T > t) = 1 - F(t), t \geq 0 \tag{1}$$

where, F(t) is cumulative distribution function, which represents the probability that a subject selected at random have a survival time less than or equal to some stated value t, given by:-

$$F(t) = P(T \leq t) = \int_0^t f(u)du, t \geq 0 \tag{2}$$

The probability density function, f(t), is given by:-

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

The hazard function is the instantaneous probability of having an event at time t (per unit time) given that one has survived (i.e. Not had an event) up to time t (Kleinbaum and Klein, 2012). It is given by:-

$$\lambda(t) = \frac{f(t)}{S(t)}F(t) = \frac{-d}{dt}lnS(t) \tag{4}$$

The cumulative hazard function is defined as:-

$$\Lambda(t) = \int_0^t \lambda(u)du = -lnS(t) \tag{5}$$

Thus;

$$S(t) = e^{-\Lambda(t)}$$

3.3.3 Cox-Proportional Hazard Model

The most frequently applied analysis method for recurrent data is the model by Andersen and Gill (1982) which is based on the common Cox proportional hazards model (Cox, 1972). The Andersen-Gill model assumes independence between all observed event times irrespective whether these event times correspond to the same/different patients. The assumption of proportional hazards is that the hazard of time-to-event at any given time for an individual in one group is proportional to the hazard at that time for an individual in the other group. Further assumptions of the Cox PH model are: the ratio of the hazard function for two individuals with different sets of covariates does not depend on time, time is measured on a continuous scale, censoring occurs randomly and uninformative censoring.

According to Kleinbaum and Klein (2005), modeling recurrent survival data can be carried out using a Cox PH model with the data layout constructed so that each subject has a line of data corresponding to each recurrent event. The model is typically use to carry out the counting Process approach is the standard Cox PH model. For recurrent survival data, a subject remains in the risk set for more than one time interval until his or her last interval, after which the subject is removed from the risk set. The hazard function for the j^{th} individual in the i^{th} group can be obtained by:-

$$\lambda_{ij}(t) = \lambda_0(t)e^{\beta' X} \tag{6}$$

where $\lambda_0(t)$ is baseline hazard, β is a p x l vector of regression coefficients and X_{ij} is the value of x for the j^{th} individual in i^{th} group i=1,..., G and j=1, ..., n_i

Checking the Assumption of Proportional Hazards

It is always a good practice to check the assumption of proportional hazards before proceeding further with other inferential activities. Schoenfeld residuals are useful to

check the proportionality of the covariates over time, that is, to check the validity of the proportional hazards assumption.

3.3.4 Frailty Models

Inference for Cox proportional hazards model (Cox, 1972) was developed under the assumption that the observations are statistically independent, at least conditionally upon covariates. However, this assumption may be violated. Thus in many epidemiological studies, failure times are clustered into groups such as families or geographical units: some unmeasured characteristics shared by the members of that cluster, such as genetic information or common environmental exposures could influence time to the studied event. In a different context, correlated data may come from recurrent events, i.e. events which occur several times within the same subject during the period of observation. In frailty models, dependence is produced by sharing an unobserved variable which is treated as a random effect, or frailty (Clayton, 1978; Petersen, Andersen and Gill, 1996).

The term frailty itself was introduced by Vaupel *et al.*, (1979) in univariate survival models and the model was substantially promoted by its application to multivariate survival data in a seminal paper by Clayton (1978) (without using the notion "frailty") on chronic disease incidence in families. Frailty models are extensions of the proportional hazards model, known as the Cox model (Cox, 1972), the most popular model in survival analysis. The frailty approach is a statistical modeling concept that aims to account for heterogeneity caused by unmeasured covariates. In statistical terms, a frailty model is a random effect model for time-to-recurrence where the random effect has a multiplicative effect on the baseline hazard function.

3.3.5 A Shared Frailty Model

A natural extension of the uni-variate frailty model will be multivariate frailty models where individuals are allowed to share the same frailty value. The assumption of a shared frailty model is that both individuals in a pair share the same frailty Z, and this is why the model is called the shared frailty model. It was introduced by Clayton (1978) and extensively studied in Therneau and Grambsch (2010), Duchateau *et al.*,

(2003). These frailties may be individual-specific or group-specific thus giving rise to the nomenclature "individual frailty" or "shared frailty" models. Shared frailty models are appropriate when you wish to model the frailties as being specific to groups of subjects, such as subjects within families. Here a shared frailty model may be used to model the degree of correlation within groups.

3.3.6 A Shared Gamma Frailty Model

The gamma distribution $\Gamma(k, \lambda)$ has been widely applied as a frailty distribution. The two parameter gamma density function is given by:-

$$f_z(Z) = \frac{k^\lambda z^{\lambda-1} e^{-kz}}{\Gamma(\lambda)} \tag{7}$$

with $\lambda > 0$ the shape parameter and k > 0 the scale parameter. The Laplace transform is

$$L(s) = \int_0^\infty e^{-zs} f_z(Z) dz = k^\lambda (s+k)^{-\lambda} \tag{8}$$

In frailty modeling the typical choice of the parameters of the gamma distribution is k= λ Using θ as notation for the variance of Z, we the have $E(Z) = 1$ and $Var(Z) =$ θ - $\frac{1}{k}$.

This distribution with parameters $(\frac{1}{\theta}, \frac{1}{\theta})$ is called a one-parameter gamma distribution with variance parameter θ. With the assumption k = λ (necessary for identify-ability reasons), the two-parameter gamma distribution turns to a one parameter distribution $\Gamma(\frac{1}{\theta}, \frac{1}{\theta})$. The functional form of the one parameter gamma distribution is given by:

$$f_z(z) = \frac{z^{\frac{1}{\theta}-1} e^{\frac{-z}{\theta}}}{\theta^{\frac{1}{\theta}} \Gamma(\frac{1}{\theta})} \tag{9}$$

where $\Gamma(.)$ is gamma function with Laplace transform that is given by:

$$L(s) = (1+\theta s)^{\frac{-1}{\theta}}, \theta > 0 \tag{10}$$

Thus, the expectation and variance of the frailty variable will be 1 and θ respectively. The shared gamma frailty model (conditional hazard) for individual j in cluster i is:-

$$\lambda_{ij}(t|X_{ij}, Z_i) = z_i \lambda_0(t) e^{x_{ij}{}'\beta} = Z_i h(t_{ij})$$

where

$$h(t_{ij}) = \lambda_0(t)e^{x_{ij}`\beta}$$

in the Cox regression model for individual j in cluster i. The Z_i are independent identically distributed following a gamma distribution, like in the uni-variate frailty models.

3.3.7 Shared Log-Normal Frailty Model

Another important frailty distribution is the log-normal distribution. The popularity of the log-normal frailty model stems mainly from the link with mixed models, where the standard assumption is that the random effects follow a normal distribution. Let W ˜ N(0,σ^2) be a normally distributed random effect and let the frailty be given by Z=e^W. The corresponding frailty has a log-normal distribution. It is function has the form:-

$$f_z(z) = \frac{1}{Z\sqrt{2\pi\sigma^2}} e^{\frac{-(log(z))^2}{2\sigma^2}}, \sigma > 0, z > 0 \tag{11}$$

Consequently, in the gamma frailty model the parameter θ denotes the variance of the frailty Z whereas in the log-normal model σ^2 denotes the variance of the random effect W = ln(Z). Both expressions can not be directly compared. Furthermore, in the log-normal model the expectation of the frailty variable is usually not one despite the fact that the expectation of the random effect W is zero. The shared log-normal frailty model has the form:-

$$\lambda_{ij}(t) = \lambda_0(t)X_{ij}`\beta + W_i \tag{12}$$

where $\lambda_{ij}(t)$ the hazard function is for the $\jmath^{th}$ individual from the $\imath^{th}$ group $\lambda_0(t)$ is the baseline hazard at time t, X_{ij} is the vector of p covariates recorded for the individual and W_i is the random effect for the $\imath^{th}$ group. In this model $\lambda_0(t)$ can be left arbitrary. The $w_i`s$, i =1,..., G are a sample (independent and identically distributed) from a density $f_w(.)$. The frailty model can be rewritten as follows:-

$$\lambda_{ij}(t) = \lambda_0(t)e^{W_i}e^{X_{ij}`\beta} = \lambda_0(t)Z_i e^{X_{ij}`\beta} \tag{13}$$

Where Z_i=e^{W_i} is known as the frailty. Model (13) is a conditional hazard function given the independent Z_i, i =1,..., G which are assumed to have a common density $f_z(.)$.

3.3.8 Penalized Likelihood Approach

Semi-parametric hazard models without frailty terms are fitted by maximization of the partial likelihood (Cox, 1972). For semi-parametric frailty models, however, we need to account for the contribution of the unobserved frailty terms. In this study we focus mainly on semi-parametric frailty models, which mean that traditional maximum likelihood estimations procedures are not appropriate for parameter estimation. An appropriate estimation method could be used to fit semi-parametric frailty models that are the expectation-maximization (EM) algorithm and the penalized likelihood approach as discussed in (Therneau and Grambsch, 2000; Duchateau and Janssen, 2008).

Penalized likelihood approach used in this study because the idea of Cox partial likelihood does not carry out in a simple manner, since the integration over frailty induces a complicated form of this likelihood. In the gamma frailty model, a compact formula for the full likelihood can be obtained by integrating out the frailty Z_i from the joint likelihood (Klein *et al.*, 1992; Nielsen *et al.*, 1992). Thus, the penalized likelihood approach has the advantage that while making no parametric assumption on the hazard or intensity functions, it yields smooth estimates of these functions.

3.3.9 Choice of Smoothing Parameter

Sometimes it is sufficient to choose the smoothing parameter heuristically, by plotting several curves and choosing the one that seems most realistic. An empirical estimate of the smoothing parameter can be provided or the smoothing parameter can be chosen by maximizing cross-validation as in (Joly *et al.*, 1999).

$$\overline{CV(k)} = \frac{1}{n} l_j(\hat{\eta}(k)) - \frac{1}{n}[\hat{I}(\hat{\eta}) + 2k\Omega]^{-1}\hat{I}(\hat{\eta}) \tag{14}$$

where:- l_j is the log-likelihood contribution of individual j. In this study, the goodness of fit of the Cox and frailty models is provided by an approximate likelihood cross-validation criterion (LCV) (Gray, 1992). Likelihood cross-validation criterion (LCV) is approximately equivalent to Akaike's criterion. Lower values of LCV indicate a better fitting model.

3.3.10 Computational Procedure (Algorithm)

The estimated parameters for the models we employed were obtained by the robust Marquardt algorithm (Marquardt, 1963) which is a combination of the Newton-Raphson algorithm and steepest descent algorithms. It is more stable than the Newton-Raphson algorithm but preserves its fast convergence property near the maximum. The iteration stops when the difference between two consecutive log-likelihoods is small, the coefficients are stable and the gradient is small enough.

3.3.11 Assessing Model Adequacy

Regardless of which type of model is fitted and how the variables are selected to be in the model, it is important to evaluate how well the model fits the data. A survival model is adequate if it represents the survival patterns in the data to an acceptable degree. This aspect of a model is known as goodness of fit. Residuals are a useful method for checking the fit of a statistical model. Residuals are central to the evaluation of model adequacy in any setting. Cox-Snell residual is the most widely used residual in the analysis of survival data (Cox and Snell, 1968). The Cox-Snell residual for the i^{th} individual is given by:

$$r_{ci} = e^{\beta' X_{ij}} \hat{H}_0(t_i) = \hat{H}_i(t_i) = -log\hat{S}_0(t_i) \tag{15}$$

where $\hat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time of that individual, $\hat{H}_i(t_i)$ and $\hat{S}_i(t_i)$ are the estimated values of the cumulative hazard and survivor functions of the i^{th} individual at t_i. The hazard function follows approximately 45 degree line at which plot depicts for $-log\hat{S}_0(t_i)$ vs time a straight line and/or for plot $log(-log\hat{S}_0(t_i))$. Log (time) a straight line through the origin with slope =1, we can say the model fit the data better.

3.4 Ethical Consideration

The research ethics review board of Jimma University would provide an ethical clearance for the study. The data has been collected after written permeation is obtained from oncology department of TASH and department of statistics write an official co-operation letter to the Hospital for the permeation. The study conducted without

informed consent since retrospective study design has been applied. Confidentiality of any information related to the patients and their clinical history has been maintained by keeping both the hard-copy and soft-copy of every collected data in a locked cabinet and password secured computer. Only the researcher would access to the de-identified data that has been kept in a secure place. All data has been coded with numbers and hospital numbers and without personal identifiers. All analysis has been on de-identified and coded data. During the study, there is no contact between the patients and the researcher. The study is noninvasive and without any harm to the patients. Then, the data obtained from the hospital has been secured.

Statistical Software Used

The statistical software used was:-

- R version 3.6.1 used for data analysis.

4 Results and Discussion

4.1 Descriptive Analysis

In this study, 421 women's who followed breast cancer treatment in TASH between 1^{st} January 2013 and 30^{th} January 2019 were considered. The response was time to recurrence of breast cancer. From the total of 997 recurrent observation, about 609 (61.1%) of them experienced recurrence of breast cancer and the remaining 388 (38.9%) were censored.

From the total of the patients 42(10%) have taken surgery of which 6(1.4%) were events. Of this total patients 153(36.3%) were those who took chemotherapy from this 77(18.3%) of them were events. And also, about 40(9.5%) breast cancer patients treated in the hospital with radiotherapy out of this 14(3.3%) of them experienced the events. About 35(8.3%) recurrent observation treated with a hormone therapy out of this 13(3.1%) of them experienced the events and 151(35.9%)patients in the hospital treated with a combination of two or more treatments out of this 56(13.3%) of them experienced the events.

Similarly, smoker incorporates 68(16.2%) of the total patients where the events experienced in 42(10%). Looking for tumor size of women with recurrent events, about 26(6.2%), 60(14.3%) and 80(19%) have a tumor size of $\leq 2cm$, 3-5 cm and $> 5cm$ experienced the events respectively, this indicate the recurrence of women with breast cancer seems higher for higher tumor size. Most of the patients come to the TASH at stage III and stage IV which is advanced stage of breast cancer and of all patients 143(34%) and 146(34.7%) stage III and IV patients, 89(21.1%) and 110(26.1%) of them were events respectively. Almost half of patients, about 199(47.3%))use oral contraceptive out of this the events experienced on 94(22.3%). Finally from the total of patients 318(75.5%) of them were married and from this 122(29%) of them experience the events. The minimum and maximum times of recurrence of women breast cancer were 5 and 32 months respectively. The median recurrence time of the breast cancer patients was 10 months in the recurrent events of women breast cancer.

Table 4.1: Frequency distribution of independent variables among the patients with breast cancer

		Recurrent Status		
Variable Names	Category	Censored	Events	Total
treatment taken	Surgery	36(8.6%))	6(1.4%)	42(10%)
	Chemotherapy	76(18.1%)	77(18.3%)	153(36.3%)
	Radiation	26(6.2%)	14(3.3%)	40(9.5%)
	Hormonal-therapy	22(5.2%)	13(3.1%)	35(8.3%)
	Combination of $\geq$2	95(22.6%)	56(13.3%)	151(35.9%)
Age	$\leq$ 35	55(13.1%)	51(12.1%)	106(25.2%)
	36-49	88(20.9%)	90(21.4%)	178(42.3%)
	$\geq$50	112(26.6%)	25(5.9%)	137(32.5%)
Smoking	non-smoker	229(54.4%)	124(29.5%)	253(83.8%)
	Smoker	26(6.2%)	42(10%)	68(16.2%)
Stage	I	18(4.3%)	7(1.7%)	25(5.9%)
	II	58(13.8%)	49(11.6%)	107(25.4%)
	III	54(12.8%)	89(21.1%)	143(34%)
	IV	36(8.6%)	110(26.1%)	146(34.7%)
Tumor size	$\leq$ 2cm	119(28.3%)	26(6.2%)	145(34.4%)
	3cm-5cm	95(22.6%)	60(14.3%)	155(36.8%)
	>5cm	41(9.7%)	80(19%)	121(28.7%)
Obesity	Normal	62(14.7%)	41(9.7%)	103(24.5%)
	Underweight	74(17.6%)	62(14.7%)	136(32.3%)
	Overweight	119(28.3%)	63(15%)	4182(43.3%)

Variable Names	Category	Recurrent Status		Total
		Censored	Events	
Histology grade	Well differentiated	89(21.1%)	34(8.1%)	123(29.2%)
	moderately differentiated	95(22.6%)	67(15.9%)	162(38.5%)
	poorly differentiated	71(16.9%))	65(15.9%)	136(32.3%)
Alcohol	No	121(28.7%)	79(18.8%)	200(47.5%)
	Yes	134(31.8%)	87(20.7%)	221(52.5%)
Family history	No	128(30.4%)	85(20.2%)	213(50.6%))
	Yes	127(30.2%)	81(19.2%)	208(49.4%)
Breast feeding	No	83(19.7%)	95(22.6%)	178(42.3%)
	Yes	172(40.9%)	71(16.9%)	243(57.7%)
Residence	Urban	2161(38.2%)	113(26.8%)	274(65.1%)
	Rural	94(22.3%)	53(12.6%)	147(34.9%)
Menopausal status	pre-menopausal	135(32.1%)	74(17.6%)	209(49.6%)
	post-menopausal	120(28.5%)	92(21.9%)	212(50.4%)
Oral contraceptive	Not used	150(35.6%)	72(17.1%)	222(52.7%)
	Used	105(24.9%)	94(22.3%)	199(47.3%)
Marital status	Single	323(5.5%)	17(4%)	40(9.5%)
	Married	196(46.6%)	122(29%)	318(75.5%)
	Widowed	22(5.2%)	18(4.3%)	40(9.5%)
	Divorced	14(3.3%)	9(2.1%)	23(5.5%)
	Total	255(60.6%)	166(39.4%)	421(100%)

4.2 Cox Proportional Hazard Model

To determine the biological, clinical and socio demography covariates which are associated with the observed time to recurrent event of breast cancer patients, first fitted Cox proportional hazard model for each risk factor before proceeding to more complicated models. Result of the uni-variable Cox proportional hazard regression model is shown in the appendix-1 (Table 5.1). Variables with p-value less than or equal to 25% in the uni-variable analysis were considered for multivariable model (Hosmer *et al.*, 2008; Bursac *et al.*, 2008). The full multivariable Cox proportional hazard model was fitted including all the potential covariates that were significant at 25% at the uni-variate level. For multivariable analysis, variables with P-value less than or equal to 5% were selected as significant covariates.

The result from the standard Cox PH model is presented on table 4.2 below. It is observed that age, stage, tumor size, histology grade, breast feeding and oral contraceptive were significantly associated with time to recurrence of breast cancer patients. The Standard Cox PH model considers different line of data contributed by the same subjects as independent contributions from different subjects.

Table 4.2: Parameter Estimates of Cox-PH Model using penalized likelihood

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value	95% CI
Age	36-49	-0.0022	0.9978	0.0979	0.98225	[0.82 , 1.21]
	$\geq$50	-0.3117	0.7322	0.1175	0.0080	[0.58 , 0.92]
Stage	II	0.2976	1.3466	0.1123	0.0081	[1.08 , 1.68]
	III	0.2865	1.3318	0.1077	0.0078	[1.08 , 1.64]
	IV	0.6330	1.8833	0.1586	<0.0001	[1.38 , 2.57]
Tumor size	3cm-5cm	0.4786	1.6138	0.1197	<0.0001	[1.28 , 2.04]
	>5cm	0.6448	1.9055	0.1215	<0.0001	[1.50 , 2.42]

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value	95% CI
Histology grade	II	0.1455	1.1566	0.1134	0.1993	[0.93 , 1.44]
	III	0.2364	1.2667	0.1117	0.0342	[1.02 , 1.58]
Breast feeding	Yes	-0.3301	0.7188	0.0859	0.00012	[0.61 , 0.85]
Oral Contraceptive	Used	0.2083	1.2316	0.0850	0.0142	[1.04 , 1.45]

4.2.1 Checking the Assumption of Cox-PH

The PH assumption of all variables included in the model was checked using the Schoenfeld residuals as described in table 4.3 below. The results show that the covariates are not statistically significant implying that the covariates are time independent because all the p-values are greater than 5%. The test of correlation (Rho) is a relation between time and residuals. It is insignificant that indicates proportional hazards assumption is fulfilled. The overall proportionality test is also not statistically significant implying that the proportionality assumption was not violated. The scatter plots of Scaled Schoenfeld residuals in appendix also used to check PH assumption. If the PH assumption is met, Schoenfeld residuals should look horizontal since the scaled Schoenfeld residuals are independent of recurrence time.

Table 4.3: Cox PH assumption checking test statistics

Variable	Category	Rho	Chisq	p-value
Treatment taken	Chemotherapy	0.05218	-1.58611	0.2079
	Radiotherapy	0.03978	0.90082	0.3426
	Hormonal	0.01692	0.17448	0.6762
	Combination of two or more	0.05864	2.03430	0.1538
Age	36-49	-0.00549	0.00973	0.9214
	$\geq$50	0.03850	0.35782	0.5497
Smoking	Yes	-0.09713	2.00309	0.1570
Stage	II	-0.02969	0.15248	0.6962
	III	-0.05021	0.46661	0.4946
	IV	-0.07048	1.05617	0.3041

Tumor size	2cm-5cm	0.10263	3.42868	0.0641
	> 5cm	0.10073	3.79636	0.0514
Obesity	Underweight	-0.05153	0.59943	0.4388
	Overweight	-0.01558	0.05379	0.8166
Histology grade	Moderately differentiated	-0.04014	0.53773	0.4634
	poorly differentiated	-0.07704	1.91115	0.1668
Alcohol	Yes	-0.00560	0.00883	0.9251
Breast feeding	Yes	-0.05511	0.66005	0.4165
Oral contraceptive	Used	-0.07222	1.44803	0.2288
GLOBAL	-	NA	18.66570	0.4785

4.3 Shared Gamma Frailty Model

In recurrent events data, subjects may have more than one events of interest. Thus, patients with the same id are considered as correlated. An extension of the Cox model can be considered by taking into account the clustered structure of the data. Thus clustering can be considering as a random effect. Here the main interest is rather in the heterogeneity between subjects. In the shared gamma frailty model, first univariable analysis was conducted and significant variables at 25% level of significance were taken to the multiple shared gamma frailty model. Result is presented in table 5.2 of the appendix. Treatment taken, age, stage, tumor size, histology grade, breast feeding and oral contraceptive were significant covariates selected from the saturated multiple shared gamma frailty model.

Table 4.4: Parameter Estimates for Shared Gamma Frailty Model uses penalized likelihood

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value	95% CI
Treatment taken	Chemotherapy	0.3679	1.4445	0.19778	0.0629	[0.98 , 2.13]
	Radiotherapy	0.3250	1.3840	0.2365	0.1694	[0.87 , 2.20]
	Hormonal	0.1762	1.1927	0.2467	0.4752	[0.74 , 1.93]
	Combination of ≥ 2	0.4343	1.5438	0.1961	0.0268	[1.05 , 2.27]
Age	36-49	-0.0061	0.9939	0.0980	0.9506	[0.82 , 1.20]
	$\geq$50	-0.3192	0.7267	0.1178	0.0067	[0.58 , 0.92]
Stage	II	0.3104	1.3640	0.1126	0.0059	[1.09 , 1.70]
	III	0.3084	1.3612	0.1079	0.0043	[1.10 , 1.68]
	IV	0.66744	1.9492	0.1592	<0.0001	[1.43 , 2.66]
Tumor size	3cm-5cm	0.5216	1.6847	0.1195	<0.0001	[1.33 , 2.13]
	>5cm	0.6798	1.9734	0.1216	<0.0001	[1.55 , 2.50]
Histology grade	II	0.1387	1.1488	0.1135	0.2217	[0.92 , 1.44]
	III	0.2565	1.2923	0.1118	0.0218	[1.04 , 1.61]
Breast feeding	Yes	-0.3672	0.6927	0.0860	<0.0001	[0.59 , 0.82]
oral contraceptive	Used	0.2236	1.2506	0.0853	0.0088	[1.06 , 1.48]
Frailty parameter, Theta:$6.49157e{-}14$ SE (H): $2.53472e{-}08$ p = < 0.5						
penalized marginal log-likelihood = -2207.38						
Convergence criteria: parameters =1.08e-05 , likelihood = 7.1e-05, gradient = 5.63e-09						
LCV = the approximate likelihood cross-validation criterion in the semi parametric case = 2.23908						

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direct and the Hessian ("sandwich estimator") respectively; the First Category of covariates' is Reference.

Parameter estimates of the shared gamma frailty model are presented in table 4.4 above interpreted as following after controlling for others prognostic factors and

accounting for frailty. The estimated hazard rate for age implied those patients who are grouped in age interval 50 or above 50 is $\hat{HR} = e^{\hat{\beta}} = e^{-0.31921178} = 0.727$ with [95% CI: 0.58 ,0.92] and the p-value are small (p-value = 0.007) which implies that age has been increase the recurrence of the diseases. Thus, the result revealed that recurrence of breast cancer for age interval 50 or above 50 of breast cancer patients at hazard rate of about 0.727 times less than patients who were grouped to age interval $\leq$35.

Looking for stages of women with breast cancer, the estimated hazard rate of breast cancer with stage II, III and IV were 1.364 with [95% CI: 1.09, 1.70], 1.361 with [95% CI: 1.10, 1.68] and 1.950 with [95% CI: 1.43, 2.66] respectively and p-value (0.006, 0.004, <0.0001) this indicates that stage II, III and IV of breast cancer has a significant effect on the increasing recurrence of women with breast cancer and the risk of recurrence breast cancer increased by 1.364, 1.361 and 1.950 for stage II, III and IV of women with breast cancer respectively as compared to stage I.

Regarding tumor size of women breast cancer, the estimated hazard rate of breast cancer patients for tumor size 3 to 5 and above 5 centimeters are estimated to be 1.685 with [95% CI: 1.33, 2.13] and 1.973 with [95% CI: 1.55, 2.50] respectively and p-value are small (<0.0001) which implies that both tumor size of women with breast cancer has a significant effect on the increasing recurrence of breast cancer and the expected hazard rate of breast cancer patients increased by 1.685 for tumor size 3 to 5 centimeters and 1.973 for tumor size above 5 centimeters as compared to women with breast cancer of tumor size 2 or below 2 centimeters.

Moreover, for those patients having histology grade III, the estimated hazard rate is to be 1.292 with [95% CI: 1.04, 1.61] in which the expected hazard rate is 1.292 times more than women having histology grade I and p-value is small (0.022) this implies that women having histology grade III has a significant effect on the increasing recurrence of breast cancer.

By observing breast feeding of women with breast cancer, the estimated hazard rate of breast cancer recurrence for women in breastfeed a child is estimated to be 0.693 with [95% CI: 0.59, 0.82] in which the expected hazard rate is 0.693 times less than women did not breastfeed child and p-value is small (<0.0001) which implies that women with breast feeding were significantly decreases the risk of recurrence of breast

cancer.

Finally, observing for women oral contraceptive use, the estimated hazard rate of breast cancer patients using oral contraceptive is 1.251 with [95% CI: 1.06, 1.48] which indicate the expected hazard rate is 1.251 times more than women did not use oral contraceptive and the p-value is small (0.0088) this indicated that a women using oral contraceptive were significantly increases the risk of recurrence of breast cancer.

4.4 Shared Log-Normal Frailty Model

Similarly, we conducted uni-variable analysis for the shared log-normal frailty model. The result of uni-variable analysis indicate treatment taken, age, smoking habit, stage, tumor size, histology grade, breast feeding and oral contraceptive were statistically significant at 25% level of significance as presented in the appendix-1 (Table 5.3). The result from the shared log-normal frailty model is presented on table 4.5 below. It is observed that age, stage, tumor size, histology grade, breast feeding and oral contraceptive were the only significant covariates selected from the saturated multiple shared log-normal frailty model.

Table 4.5: Parameter Estimates of Shared Log-normal Frailty Model using penalized likelihood

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value	95% CI
Age	36-49	-0.0124	0.9877	0.1052	0.9062	[0.80 , 1.21]
	$\geq$50	-0.3346	0.7156	0.1245	0.0072	[0.56 , 0.91]
Stage	II	0.3428	1.4090	0.1187	0.0039	[1.12 , 1.78]
	III	0.3137	1.3685	0.1126	0.0053	[1.10 , 1.71]
	IV	0.7286	2.0722	0.1696	<0.0001	[1.49 , 2.89]
Tumor size	3cm-5cm	0.4598	1.5838	0.1234	0.00019	[1.24 , 2.02]
	>5cm	0.6640	1.9425	0.1253	<0.0001	[1.52 , 2.48]

Variable names	category	$\hat{\beta}$	exp($\hat{\beta}$)	SE coef (H)	p-value	95% CI
Histology grade	II	0.1160	1.1231	0.1183	0.3266	[0.89 , 1.42]
	III	0.2492	1.2831	0.1162	0.0319	[1.02 , 1.61]
Breast feeding	Yes	-0.3514	0.7037	0.0894	<0.0001	[0.59 , 0.84]
oral contraceptive	Used	0.2233	1.2503	0.0900	0.0130	[1.05 , 1.49]

Frailty parameter, Sigma Square:0.075618 SE (H): 0.00864336 p = $< e^{-16}$

penalized marginal log-likelihood = -2162.87

Convergence criteria: parameters =5.95e-07, likelihood = 0.000894, gradient = 6.81e-09

LCV = the approximate likelihood cross-validation criterion in the semi parametrical case = 2.19644

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direct and the Hessian ("sandwich estimator") respectively; the First Category of covariates' is Reference.

Parameter estimates of the shared log-normal frailty model are presented in table 4.5 above are interpreted as following after controlling for others prognostic factors and accounting for frailty. The estimated hazard rate for age implied those patients who are grouped in age interval 50 or above 50 is $\hat{HR} = e^{\hat{\beta}} = e^{-0.3345833} = 0.716$) with [95% CI: 0.56- 0.91] and p-value are small (p-value= 0.0072) which implies that younger women, particularly those under the age of 35, have a higher risk of recurrence of breast cancer and the result revealed that age interval 50 or above 50 of breast cancer patients has rate of recurrence 0.716 times less than patients who were grouped to age interval $\leq$35.

Looking for stages of women with breast cancer, the estimated hazard rate of breast cancer for stage II, III and IV were 1.409 with [95% CI: 1.12 - 1.78], 1.368 with [95% CI: 1.10 - 1.71] and 2.0722 with [95% CI: 1.49 - 2.89] and the p-value are small (0.0039, 0.0053, <0.0001) respectively and this indicates that women with advanced stages have a greater risk for recurrence of breast cancer. Thus, the expected hazard rate of women with recurrence of breast cancer increased by 1.409, 1.368 and 2.0722 for stage II, III

and IV respectively as compared to stage I.

Regarding tumor size of women breast cancer, the estimated hazard rate of recurrence for tumor size 3 to 5 and above 5 centimeters are estimated to be 1.584 with [95% CI: 1.24 - 2.02] and 1.942 with [95% CI: 1.52 - 2.48] respectively and p-value are small (<0.0001, <0.0001) which implies that both tumor size of women with breast cancer has a significant effect on the increasing recurrence of women breast cancer and women with larger tumors have a greater risk of recurrent breast cancer and also the expected hazard rate of women with breast cancer increased by 1.584 for tumor size 3 to 5 centimeters and 1.942 for tumor size above 5 centimeters as compared to tumor size 2 or below 2 centimeters.

Moreover, for those patients having histology grade three, the estimated hazard rate is estimated to be 1.283 with [95% CI: 1.02 - 1.61] in which the expected hazard rate is 1.283 times more than women having histology grade one and the p-value is small (0.032) which implies that women having histology grade three has a significant effect on the increasing recurrence of breast cancer.

By observing breast feeding of women with breast cancer, the estimated hazard rate of breast cancer recurrence for women in breastfeed a child is estimated to be 0.7037 with [95% CI: 0.59 - 0.84] which is the expected hazard rate is 0.7037 times less than women did not breastfeed child and the p-value is small (<0.0001) which implies that women breast feeding a child is significantly decreases the risk of recurrence of breast cancer.

Finally, observing for women oral contraceptive use, the estimated hazard rate of breast cancer patients using oral contraceptive is estimated to be 1.250 with [95% CI: 1.05 - 1.49] which implies the expected hazard rate is 1.250 times more than women did not use oral contraceptive and the p-value is small (0.013) this implies that women use oral contraceptive is significantly increases the risk of recurrence of breast cancer.

Test hypothesis for the variance term of shared log-normal frailty term is given by:

$$H_0{:}\sigma^2 = 0 \text{ vs } H_1{:}\ \sigma^2 > 0$$

variance of frailty term (Sigma Square): σ^2= 0.0756 (SE(H):0.0086) is significantly different from zero, meaning that there is heterogeneity between the subjects explained

by non-observed covariates. We can deduce this by using a modified Wald test:$W_m(\sigma^2)$ = 0.075618 /0.00864336=8.75, with the critical value for a normal one-sided test. The modified Wald test (W_m) is a significance test for the variance of the random effects distribution occurring on the boundary of the parameter space. The usual squared Wald statistic is simplified to a mixture of two distributions and hence the critical values must be derived from this mixture (Molenberghs and Verbeke, 2007). In the case our result we have a p-value that is less than 5% for shared log-normal frailty but not for shared gamma frailty distribution. This mean that there is a significant frailty effect, that within subject correlation cannot be ignored for shared log-normal frailty but not for shared gamma frailty.

4.5 Assessment of Model Adequacy

From the plot of Cox-Snell residuals against the cumulative hazard given in the figure below, if the model fits the data, the plot of cumulative hazard function against Cox-Snell residuals should be approximately a straight line with slope one. Thus, the Cox-Snell residuals indicated that shared log-normal frailty model fit the data better.

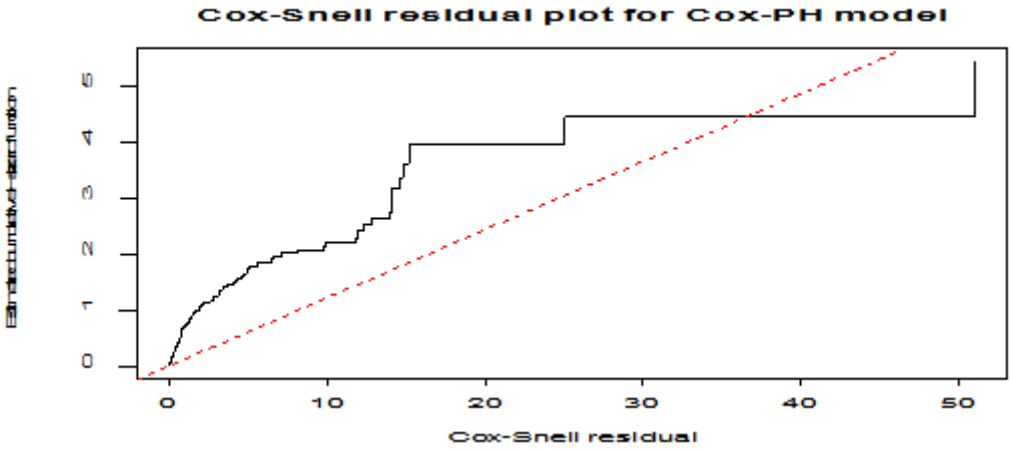

Figure 4.1: Cox-Snell residuals plot for cox-Ph model

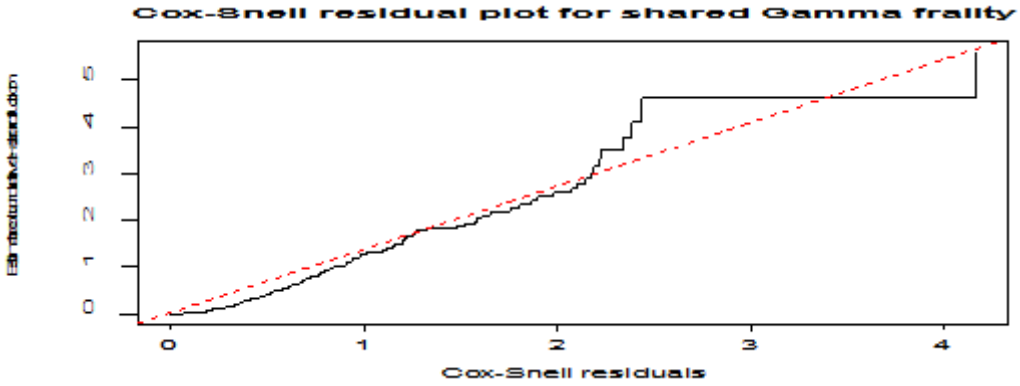

Figure 4.2: Cox-Snell residuals plot for shared gamma frailty model

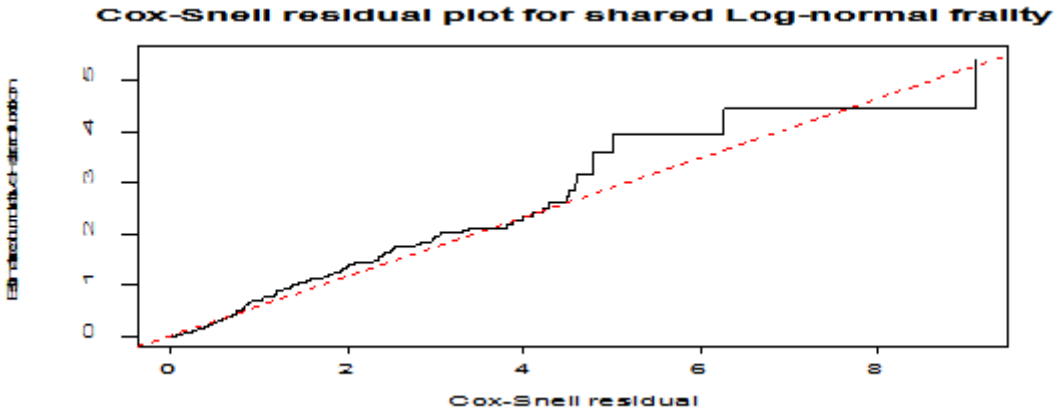

Figure 4.3: Cox-Snell residuals plot for shared log-normal frailty model

4.6 Comparison of the Cox-PH and Shared Frailty Models

Efficiency of the fitted models was compared using penalized marginal log-likelihood and LCV (likelihood cross-validation) criterion. The likelihood cross-validation criterion assesses the goodness of fit of a statistical model (Gray, 1988). In case of parametric approach, LCV is approximately equivalent to Akaike Information criterion (AIC). Lower values of LCV indicate a better fitting model. Table 4.6 depicts the LCV results of Cox PH, shared gamma frailty and shared log-normal frailty models. The shared log-normal frailty model was chosen as the best fit for our recurrent events data on breast cancer based on the residual analysis and minimum value of LCV. The Wald test results indicated that the heterogeneity parameter was significant, implying that there is a significant frailty effect, or that within subject correlation cannot be ignored.

Table 4.6: Comparison of Cox PH, Shared Gamma and Log-normal Frailty Models

Model	Frailty Parameter	SE	PMlogL	LCV
Cox-PH	–	–	-2217.487	2.24721
Shared Gamma Frailty	$6.49157\ e^{-14}$	$2.53472\ e^{-8}$	-2207.38	2.23908
Shared Log Normal Frailty	0.0756	0.0086	-2162.87	2.19644

From the above model comparison using penalized marginal log-likelihood and LCV (likelihood cross-validation) criterion. Using LCV The shared log-normal frailty model was chosen as the best fit for the recurrent events data on breast cancer based on the residual analysis and minimum value of LCV.

4.7 Discussion

The main aim of this study is to identify factors affecting the recurrence of women breast cancer, which was obtained from TASH. The most frequently applied analysis method for recurrent time-to-event data is the model by Andersen and Gill was used to analyze breast cancer data set. The Andersen-Gill model assumes independence between all observed event times (Andersen and Gill, 1982). In addition, frailty models used to account for the dependence among the recurrent event times based on Andersen-Gill (A-G) survival model (Pickels *et al.*, 1994; Yashin *et al.*, 1995).

From the total of 997 case of recurrent observation, 609 (61.1%) them experienced the event. From the total of 421 patients 42(10%),153(36.3%), 40(9.5%), 35(8.3%), 151(35.9%) breast cancer patients treated in the hospital with surgery, chemotherapy, radiotherapy, hormone-therapy and combination of two or more treatments respectively and out of this 6(1.4%), 77(18.3%), 14(3.3%), 13(3.1%), 56(13.3%) of them experienced the events respectively. The minimum and maximum times of recurrence of women breast cancer were 5 and 32 months respectively. The median recurrence time of the breast cancer patients was 10 months in the recurrent events of women breast cancer.

Efficiency of the fitted models was compared using penalized marginal log-likelihood and LCV (likelihood cross-validation) criterion and shared log-normal frailty model found to be the best fit from the Cox PH and shared gamma frailty model (Gray, 1992).

Under uni-variable analysis the shared log-normal Frailty model shows that treatments taken, age, smoking, tumor size, stages of breast cancer, obesity, histology grade, alcohol use, family history of breast cancer, breast feeding and oral contraceptives were significantly associated with recurrence of women breast cancer at 25% level of significance (Hosmer *et al.*, 2008; Bursac *et al.*, 2008).

From result of multivariable analysis of shared log-normal Frailty model the recurrence of women breast cancer were significantly affected by age, tumor size, stages of breast cancer, histology grade, breast feeding and oral contraceptives.

From the result of this study the younger women experienced the greatest hazard rate for recurrence of breast cancer. The study by Dignam in 2009 shows the same

results. In addition, the histology grade at diagnosis was significantly affected the recurrence of breast cancer and the hazard rate was high for women in histology grade III as compared to women in histology grade I. This was also indicated by study done (Dignam *et al.*, 2009; Mauguen *et al.*, 2013).

Similarly the hazard rate of women with breast cancer for tumor size 3 to 5 centimeters and above 5 centimeters were increase as compared to women with breast cancer of tumor size 2 or below 2 centimeters this implies that tumor size has a significant effect on increasing the risk of breast cancer recurrences. This study was justified by study by Mauguen *et al.*, (2013) and Rondeau *et al.*, (2007).

The stages of breast cancer have significant effect on the recurrence of women breast cancer. The study done by Demicheli in 2010 and Dignam in 2009 also shows that the stages of breast cancer at diagnosis have been significantly affect the recurrence of women breast cancer. From the results of these two studies the hazard rate of recurrence of breast cancer was greatest as the stage increases(Demicheli *et al.*, 2010 and Dignam *et al.*, 2009).

Oral contraceptives were associated with increasing recurrence of women breast cancer. This implies that oral contraceptives facade a higher risk of breast cancer recurrence. The current study was consistent with the study ((Lu, *et al.*, 2011; Saxe, *et al.*, 1999). Furthermore, Women treated for breast cancer that previously breast fed their babies have lower risk of recurrence than those who did not. These findings are consistent with a study conducted by (Anderson *et al.*, 2014).

The model adequacy of shared log-normal frailty model checked by plotting Cox-Snell residuals against cumulative hazard function of residuals and it's result shows that the log-normal frailty model fit women breast cancer data well, since the plot makes approximately straight lines through the origin for women breast cancer data set.

5 Conclusion and Recommendation

5.1 Conclusion

The data consisted of 997 observations from a total of 421 patients. About (61.1%) of them experienced the event of interest while the remaining (38.9%) of observations did not experience the event of interest throughout the study period.

The main goal of this study was to investigate determinants of recurrence of women breast cancer and to fit best model for recurrence of breast cancer. The standard Cox PH, shared gamma frailty and shared log-normal Frailty models were fitted. The value of the LCV was used to identify the best model. Accordingly, the shared log-normal frailty models provide suitable choice for the recurrent event time of the breast cancer as compared to standard Cox proportional hazard and shared gamma frailty models.

From the result of shared log-normal frailty model age, stage, tumor size, histology grade, breast feeding and oral contraceptive were found to be statistically significant factors for recurrence of women breast cancer. Of all this significant covariates stage (II, III, IV), tumor size ((3-5)cm, >5 cm), histology grade (poorly differentiated) and oral contraceptive (use) were significantly increases the risk of recurrence of women breast cancer. While, breast feeding was significantly decreases the risk of recurrence of women breast cancer.

5.2 Recommendations

Based on the study finding the following recommendations are forwarded:-

- Our result showed age, stage, tumor size, histology grade, breast feeding and oral contraceptive were statistically associated with recurrence of breast cancer. This calls for actions on improvement of patients health status based on these significant risk factors.

- The ministry of health of the country, policy makers and Tikur Anbessa Specialized Hospital should work on awareness of the disease so that the women has been protect themselves from the diseases by being treated at early stage of the

disease because of the disease is curable.

- The physicians are expected to record additional information of the patients history such as Physical activities, age at marriage, nutritional diets and etc., because these are the expected risk factors from many literature's.
- Awareness has to be given for the society regarding oral contraceptive and breast feeding. The mass media can play an effective role in this regard.
- Further studies considering other methods of recurrent events such as marginal, conditional for the calendar time scale, gap time scale and also joint shared frailty model for recurrent events and terminal event (death) are recommended.

5.3 Limitations of the Study

Some of the limitations of the study are:-

- The study was conducted based on secondary data which might have incomplete and biased information.
- Even if the study is hospital based retrospective study, on follow-up it is often a physical and financial burden for them to return for follow-up, so the study may have some limitation in covering all patients that take the first treatment.
- As the data is gathered from patients card the study has limited number of variables considered as risk factors for the recurrence of breast cancer.

References

Abate S, Yilma Z, Assefa M, Tigeneh W (2016) Trends of Breast Cancer in Ethiopia. *Int J Cancer Res Mol Mech **2(1)***: doi http://dx.doi. org/10.16966/2381-3318.121

Adams S, Schmid P, Rugo H (2017) Phase 2 study of pembrolizumab (pembro) monotherapy for previously treated metastatic triple-negative breast cancer.

Adesina A, Chumba D, Nelson AM, Orem J, Roberts D, Wabinga H, *et al.*, (2013) 'Improvement of pathology in sub-Saharan Africa'. The Lancet Oncology. 14(4)

Ahmedin Jemal, DVM P, Freddie Bray (2012) Cancer Burden in Africa and Opportu nities for Prevention. *Cancer 118:4372-84.*

Aalen(1980) Non-parametric estimation of partial transition probabilities in multiple decrements models. *Ann Stat.*

American Cancer society (2015) *Breast cancer facts and figures 2015-2016.* Atlanta, Georgia: American cancer society; 2015.

American Cancer Society (2016) *Breast Cancer Facts & Figures 2016-2017.* Atlanta: Inc. 2016.

American Cancer Society (2017) *Breast Cancer Facts & Figures 2017-2018.* Atlanta: Inc. 2017.

American Cancer Society (2018) *Cancer Facts & Figures.* Atlanta: Inc.2018

American Joint Committee on Breast Cancer Staging (2016) 7th Edition.

Andersen P, Gill R (1982) Cox regression model for counting processes: A large

sample study. *Ann Stat;* ***10(4)****:1100-20.*

Andersen P, Borgan , Gill R, Keiding N: Statistical Models Based on Counting Processes. 1993, Springer: New York

Anderson KN, Schwab RB, Martinez ME (2014) 'Reproductive risk factors and breast cancer subtypes: a review of the literature'. Breast Cancer Res Treat. **144(1)**:1-10. [PMC free article] [PubMed] [Google Scholar]

Baghestani A., Moghaddam S., Majd H., Akbari M, Nafissi N, Gohari K *et al.*, (2016) Survival Analysis of Patients with Breast Cancer using Weibull Parametric Model. *Asian Pac J Cancer Prev* ***16(18)***:8567- 8571.

Brinton L, Figueroa J, Awuah B, Yarney J, Wiafe S, Wood S, *et al.* (2014) 'Breast cancer in Sub-Saharan Africa: opportunities for prevention'. Breast cancer research and treatment.144(3):467-78.

Britt K, Ashworth A, Smalley M (2007) 'Pregnancy and the risk of breast cancer'. End ocrine-related Cancer.14: 907-933.

Bursac Z, Gauss C, Williams D, Hosmer D.(2008) Purposeful selection of variables in logistic regression. *Source Code Biol Med. 3:17.* doi 10.1186/1751-0473-3-17

Cardoso F, Harbeck N, Fallowfield L, Kyriakides S, Senkus E(2012) Locally recurrent or metastatic breast cancer. Ann Oncol. 2012;23 Suppl 7:vii119. doi: 10.1093/annonc/mds232.

Chen W., Rosner B., Hankinson S., Colditz G., & Willett W (2011) 'Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk'. JAMA. 306: 1884-1890.

Cheng L, Swartz MD, Zhao H, Kapadia AS, Lai D, Rowan PJ, *et al* (2012). Haz ard of recurrence among women after primary breast cancer treatment a 10-year follow-up using data from SEER Medicare. *Cancer Epidemiol Biomarkers Prev;* 21:8009.

Clayton, D. (1978) Model for Association in Bi-variate Life Tables and Its Applica tion in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence.

Cook, R. J. and Lawless, J. F. (2007). The Statistical Analysis of Recurrent Events, Statistics for Biology and Health, Springer.

Cox DR(1972) Regression models and life-tables. J R Stat Soc Series B Method ological. **34(2)**:187-220.

Da Costa Vieira, R., Biller, G., Uemura, G., Ruiz, C., & Curado, M. (2017) Breast cancer screening in developing countries. *Clinics(Sao Paulo, Brazil),* ***72(4)***, 244-253. doi:10.6061/clinics/2017(04)09.

David G. Kleinbaum and Mitchel Klein (2005): Survival Analysis, Statistics for Bi ology and Health. Second Edition, Spring Street, New York, NY, 10013, USA,

Demicheli, R., Ardoino, I., Boracchi, P. *et al.*(2010) 'Recurrence and mortality ac cording to Estrogen Receptor status for breast cancer patients undergoing cons ervative surgery'. Ipsilateral breast tumour recurrence dynamics provides clues for tumour biology within the residual breast. *BMC Cancer 10, 656 (2010)* doi:10.1186/ 1471-2407-10-656

Dignam J, Dukic V, Anderson S, Mamounas E, Wickerham D, Wolmark N (2009) 'Hazard of recurrence and adjutant treatment effects over time in lymph node-negative breast cancer'. Breast Cancer Res Treat 116:595-602. doi:10.1016/

j.mcm.2010.02.002.

Duchateau, L., Janssen P., Kezic I., and Fort-pied C.(2003). Evolution of recurrent asthma event rate over time in frailty models. Applied Statistics 52, 355-363.

Duchateau, L. and P. Janssen (2008). The Frailty Model. New York: Springer-Ver lag.

Faupel-Badger J, Arcaro K, Balkam J, *et al.* (2013) Postpartum remodeling, lacta, tion and breast cancer risk: summary of a National Cancer Institute-sponsored workshop. *J Natl Cancer Inst. 2013;105: 166-174.*

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M (2015) Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBO CAN 2012. *International Journal of Cancer. 2015;**136(5)***:E359-E86.

Federal Ministry of Healthy (2015) *National cancer control plan, 2016-2020* FMOH; 2015.

Gaud M, Gapstur S, Sun J, Diver W, Hannan L, Thun M (2013) Active smoking and breast cancer risk: original cohort data and meta-analysis. *J Natl Cancer Inst.* 2013;105: 515-525.

Geurts Y, Witteveen A, Bretveld R, Poortmans P, Sonke G, Strobbe L, Sies ling S. Patterns and predictors of first and subsequent recurrence in women with early breast cancer. *Breast Cancer Res Treat. 2017;165:709-20.*

Global Burden of Disease Cancer Collaboration (2017) Fitzmaurice C, Allen C, Barber R, Barregard L, Bhutta Z, *et al.* Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: 'A systematic analysis for the Global

Burden of Disease Study'. JAMA Oncol. 2017;3:524-48.

Gray R, (1992) Flexible methods for analyzing survival data using splines, with appli cation to breast cancer prognosis. *JASA, 87 942-951.*

Hastie, T. and Tibshirani, R. (1992). *Varying coefficients models.* Journal of the Royal Statistical Society, SeriesB55, 757-796.

Hosmer D, Lemeshow S, May S (2008) Applied Survival Analysis: Regression Modeling of Time-to-Event Data. Vol. *618. New York, NY: Wiley-Interscience.*

Howlader N, Noone AM, Krapcho M, *et al*(2017). SEER Cancer Statistics Review, 1975-2014. based on November 2016 SEER data submission, posted to the SEER web site, April 2017. *Bethesda, MD: NCI, 2017.*

International Agency for Research on cancer (2013) Latest world cancer statistics Global cancer burden rises to 14.1 million new cases in 2012: Marked increase in breast cancers must be addressed. WHO: Press release N 223.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E. & Forman, D. (2011) Global cancer statistics. *A cancer journal for clinicians American cancer society,61, 69-90*

Joly P., Leteneur D. L., Alioum A., Commenges D., (1999): a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data, Computer Method and Program.

Kaplan EL. & Meier P. (1958) Non-parametric estimation from incomplete observa tions. *J am Stat Assoc 1958(53)*:457-81

Klein, J.P. (1992). Semi parametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm, 'Biometrics 48, 795-806'.

Kleinbaum and Klein,J.P. (2005). Survival Analysis: A Self- Learning Text. Sec ond Edition, Springer.

Kleinbaum, Klein . Survival Analysis: A Self-Learning Text. New York, NY: Springer Publishing Company; 2012.

La Vecchia C, Giordano SH, Hortobagyi GN, Chabner B. Overweight, obesity, dia betes, and risk of breast cancer: interlocking pieces of the puzzle. Oncologist. 2011;16: 726-729.

Liu Y, Nguyen N, Colditz GA.(2015) Links between alcohol consumption and breast cancer: a look at the evidence. *Womans Health (London, England). 2015;11*: 65-77

Lu Y, Ma H, Malone KE, Norman SA, Sullivan-Halley J, *et al.*,(2011) Oral contra ceptive use and survival in women with invasive breast cancer. 'Cancer Epidemiol Biomarkers' Prev 20: 1391-1397

Mauguen A, Rachet B, Mathoulin-Plissier S, *et al*(2013) Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Stat Med. 2013;**32(30)**:5366-80.*

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear param eters. *Journal of the society for Industrial and Applied Mathematics **11 (2)***, 431-441

Nielsen G., Gill R., Andersen P. and Sorensen T, (1992). A Counting Process Ap proach to Maximum Likelihood Estimation in Frailty Models," Scandinavian *Jour nal of Statistics 19, 25-43.*

Parkin DM, Bray F, Ferlay J, Jemal A (2012) 'Cancer in Africa'. Cancer Epidemiology

and Prevention Biomarkers 2014, 23(6):953-966.

Petersen, P. K. Andersen, and R. D. Gill, "Variance components models for survival data," Statistica Neerlandica vol. 50 pp. 193211, 1996.

Pickels, A., Crouchley, R., Simonoff, E.L., Meyer, J., Rutter, M., Hewitt, J. and Silbery J., (1994) *Survival models for developmental genetic data: Age at onset of puberty and antisocial behavior in twins.* Genetic Epidemiology, 11: 155-170.

Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P(2007) Joint frailty models for recurring events and death using maximum penalized like lihood estimation: Application on cancer events. *Biostatistices. 2007;8(4):708-21.*

Rondeau V., (2010)*Statistical models for recurrent events and death: Application to cancer events.* Mathematical and Computer Modeling (2010) 949-955. France University Victor Segalen Bordeaux2, INSERM, CR897 (Biostatistics).

S. Moody, D. Perez, T. Pan (2005) 'The transcriptional repressor Snail promotes mam mary tumor recurrence'. Cancer Cell, vol. 8, no. 3, pp. 197-209, 2005.

Saxe G, Rock C, Wicha M Schottenfeld D (1999) *Diet and risk for breast cancer recurrence and survival.* Breast Cancer Res Treat 53: 241-253.

Tefera,B., Assefa, M., Abebe, B. & Rauch, D. (2016) Patterns of Cancer in University of Gondar Hospital: North-West Ethiopia. *Journal of oncology medicine*

Therneau T, Grambsch P. (2000) *Modeling survival data.* New York: Springer

Therneau TM, Grambsch PM.(2010) *Modeling Survival Data: Extending the Cox Model.* New York, NY: Springer-Verlag

Ullah S, Gabbett TJ, Finch CF(2014) *Statistical modeling for recurrent events: an*

application to sports injuries. Br J Sports Med. 2014;48:1287-93.

Vaupel J., Manton K., and Stallard E. (1979). *The impact of heterogeneity in in dividual frailty on the dynamics of mortality.* Demography 16.3, pp. 439-454.

Wang YC, Meyerson L, Tang Y, Qian N(2009) Statistical methods for the analysis of relapse data in ms clinical trials. *J Neurolog Sci. 2009;**285(1-2)**:206-11.*

World Health Organization (2015) Cancer a Growing Public Health Concern for Ethiopia. Available at www.who.int

Yashin, A., Vaupel, J., and Jachine, I. (1995) Correlated individual Frailty: An ad vantageous approach to survival analysis of bivariate data. *Mathematical population studies, 5:145-149.*

Appendices

Appendix-1:-Some selective relevant summary tables and graphs

Table 5.1: Parameter Estimation of univariable Coxph Model

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value
Treatment taken	Chemotherapy	0.4729	1.6047	0.1945	0.0150
	Radiotherapy	0.4121	1.5100	0.2330	0.0770
	Hormonal	0.1116	1.1180	0.2442	0.6480
	Combination of ≥ 2	0.5072	1.6606	0.1929	0.0086
Age	36-49	0.0065	1.0065	0.0966	0.9462
	$\geq$50	-0.3833	0.6816	0.1150	0.0009
Smoking	Yes	0.2471	1.2803	0.1146	0.0310
Residence	rural	-0.0934	0.9108	0.0857	0.2760
Number of child	3-5	-0.0396	0.9612	0.0915	0.6652
	5	0.0115	1.0116	0.1094	0.9161
Stage	II	0.4926	1.6366	0.1107	0.0001
	III	0.3622	1.43644	0.1064	0.0007
	IV	0.7616	2.1417	0.1577	0.0001
Tumor size	3cm-5cm	0.4569	1.5791	0.1187	0.0001
	>5cm	0.8237	2.2790	0.1178	0.0001
Obesity	underweight	-0.0236	0.9766	0.1084	0.8274
	overweight	-0.0117	0.9884	0.1032	0.9098

Variable names	category	$\hat{\beta}$	exp($\hat{\beta}$)	SE coef (H)	p-value
Histology grade	II	0.2041	1.2264	0.1120	0.0685
	III	0.3379	1.4020	0.1102	0.0022
Alcohol consumption	Yes	0.0360	1.0368	0.0811	0.6565
Family history	With	-0.0290	0.9714	0.0812	0.7208
Breast feeding	Yes	-0.3744	0.6877	0.0842	0.0001
Menopausal status	postmenopausal	-0.0117	0.9884	0.0811	0.8856
Oral Contraceptive	Used	0.2477	1.2811	0.08254	0.0027

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direct and the Hessian ("sandwich estimator") respectively

Table 5.2: Parameter Estimates of uni-variable Shared Gamma Frailty Model

Variable names	category	$\hat{\beta}$	exp($\hat{\beta}$)	SE coef (H)	p-value
Treatment taken	Chemotherapy	0.4155	1.5152	0.1944	0.0326
	Radiotherapy	0.3641	1.4392	0.2327	0.1177
	Hormonal	0.1728	1.1887	0.2439	0.4786
	Combination of ≥ 2	0.4960	1.6423	0.1934	0.0103
Age	36-49	0.0065	1.0066	0.0966	0.9462
	$\geq$50	-0.3833	0.6816	0.1149	0.0008
Smoking	Yes	0.2471	1.2803	0.1146	0.0311
Residence	rural	-0.0934	0.9108	0.0856	0.2760
Number of child	3-5	-0.0396	0.9612	0.0916	0.6652
	5	0.0115	1.0116	0.1094	0.9161
Breast feeding	Yes	-0.3744	0.6877	0.0843	<0.0001

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value
Stage	II	0.4927	1.6368	0.1107	<0.0001
	III	0.3622	1.4366	0.1064	0.0007
	IV	0.7617	2.1419	0.1577	<0.0001
Tumor size	3cm-5cm	0.4569	1.5791	0.1187	<0.0001
	>5cm	0.8238	2.2790	0.1177	<0.0001
Obesity	underweight	-0.0236	0.9766	0.1084	0.8274
	overweight	-0.0117	0.9884	0.1033	0.9098
Histology grade	II	0.2041	1.2264	0.1120	0.0685
	III	0.3378	1.4019	0.1102	0.0022
Alcohol consumption	Yes	0.0361	1.0368	0.0812	0.6565
Family history	With	-0.0290	0.9714	0.0812	0.7208
Menopausal status	postmenopausal	-0.0116	0.9885	0.0812	0.8868
Oral Contraceptive	Used	0.2477	1.2810	0.0825	0.0027

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direct and the Hessian ("sandwich estimator") respectively.

Table 5.3: Parameter Estimates of uni-variable Shared Log-normal Frailty Model.

Variable names	category	$\hat{\beta}$	$\exp(\hat{\beta})$	SE coef (H)	p-value
Treatment taken	Chemotherapy	0.5051	1.6571	0.1977	0.0106
	Radiotherapy	0.4386	1.5505	0.2380	0.0653
	Hormonal	0.1231	1.1309	0.2493	0.6216
	Combination of ≥ 2	0.5368	1.7106	0.1961	0.0062
Breast feeding	Yes	-0.3986	0.6712	0.0879	<0.0001

Variable names	category	$\hat{\beta}$	exp($\hat{\beta}$)	SE coef (H)	p-value
Age	36-49	0.0127	1.0127	0.1031	0.9022
	≥50	-0.4013	0.6694	0.1211	0.0009
Smoking	Yes	0.2539	1.2891	0.1210	0.0360
Residence	rural	-0.0983	0.9064	0.0912	0.2811
Number of child	3-5	-0.0390	0.96178	0.0972	0.6885
	5	0.01137	1.0114	0.1167	0.9226
Stage	II	0.5182	1.6789	0.1163	<0.0001
	III	0.3850	1.4697	0.1109	0.0005
	IV	0.8157	2.2607	0.1681	<0.0001
Tumor size	3cm-5cm	0.4788	1.6141	0.1227	<0.0001
	>5cm	0.8694	2.3856	0.1218	<0.0001
Obesity	underweight	-0.0273	0.9730	0.1149	0.8120
	overweight	-0.0159	0.9842	0.1094	0.8841
Histology grade	II	0.2124	1.2366	0.1163	0.0679
	III	0.3662	1.4422	0.1145	0.0014
Alcohol consumption	Yes	0.0349	1.0355	0.0864	0.6862
Family history	With	-0.0343	0.9663	0.0862	0.6910
Menopausal status	postmenopausal	-0.0135	0.9866	0.0862	0.8752
Oral Contraceptive	Used	0.2688	1.3084	0.0867	0.0019

Coef: estimated coefficient for each covariates; exp (coef): exponentiation value of coefficient, se coef (H) and se coef (HIH): estimator of standard error direct and the Hessian ("sandwich estimator") respectively.

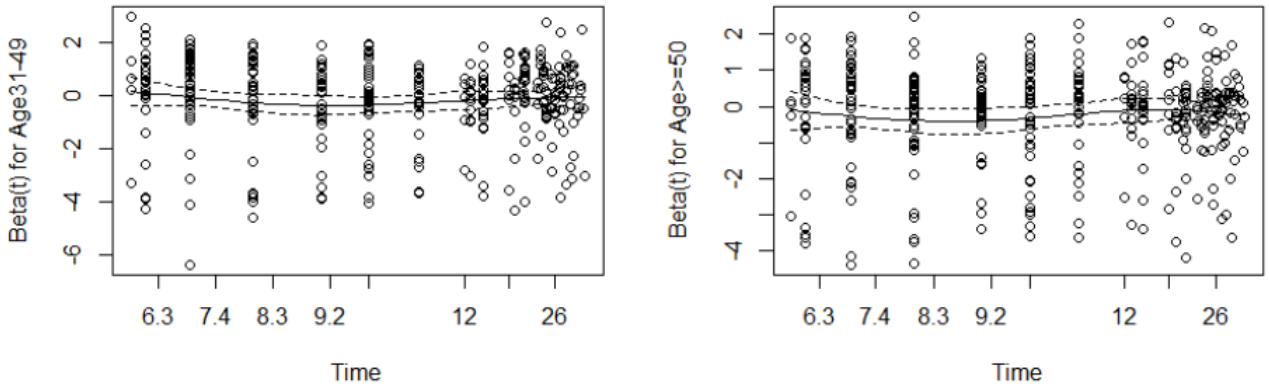

Figure 5.1: Plots of Scaled Schoenfeld Residuals for age categories

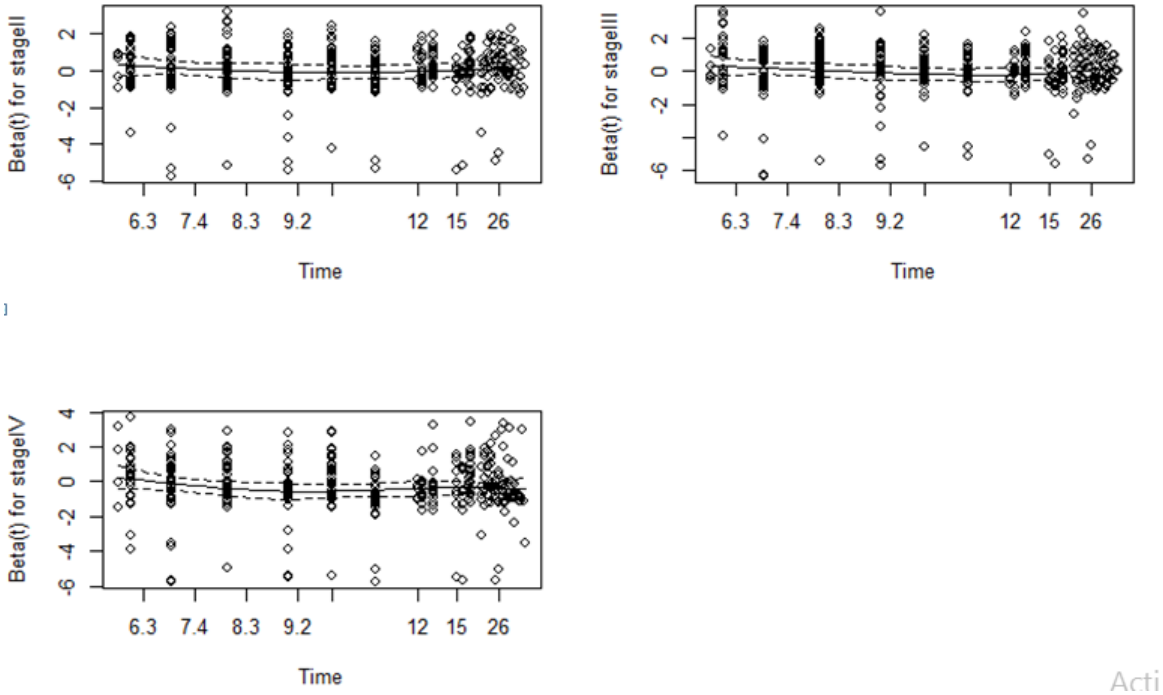

Figure 5.2: Plots of Scaled Schoenfeld Residuals for stage categories

Appendix-2:-Information Sheet

Introduction:- this information sheet is prepared for Tikur Anbessa Specialized Hospital, oncology department, Addis Ababa, Ethiopia. The aim of the form is to make clear about the purpose of thesis, data collection procedures and to get permission for data collection.

Objective:- The aim of the study is to investigate determinants of the survival of women breast cancer patients recurrence in Tikur Anbessa Specialized Hospital, Ethiopia using shared frailty models.

Data Collection Procedure:- In order to achieve the above objective, information, which is necessary for the study, will be taken from the registration log book and patients' registration card; if any inadequate information is countered it is checked from the file and excluded from analysis if proven to be inadequate. In order to come up with the above mentioned findings, total document of program clients enrolled during 1st January 2013 to 30th January 2019 will be seen and a review of the required information from the records are made by using the checklist.

Risk:- Since the study will be conducted by taking appropriate information from medical chart, it will not inflict any harm on the patients. The name or any other identifying information will not be recorded and all information taken from the chart will be kept strictly confidential and in a safe place. The information extracted will be kept secured and the information retrieved will only be used for the study purpose.

Benefits:- the thesis has no direct benefit for those whose document/ record is included in this thesis. However, indirectly the result of this study might be used to improve awareness on the factors that triggers the recurrence of breast cancer patients. It also enables to provide scientific information about the finding to Ministry of health in Ethiopia that helps policy makers to enhance the awareness of the society about factors that increase the probability of recurrence due to breast cancer which is protected and curable if it is screened and treated in its earlier stage with appropriate treatment.

Confidentiality:- To ensure confidentiality the data on the chart will be collected by those individuals who are working in oncology unit nurse and information will be collected without the name of the clients. The information collected from this thesis will be kept confidential and will be stored in a file. In addition, it will not be revealed

to anyone except the investigator and it will be kept in key and locked system with computer password.

Person to contact:- This thesis will be reviewed and approved by the institutional review board of college of Natural sciences, Jimma University.

Permission:- Lastly but not least, you are kindly requested to permit and forward your permission to concerned body in your organization so that the I can get cooperation from the data clerks and other responsible bodies in place.

Data Extraction Form

Data extraction form, for the Determinants of the recurrence of Women Breast Cancer in Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia, of women breast cancer data (Starting from 1^{st} January 2013 to 30^{th} January 2019).

1. Patient Age at diagnosis (in year): ◯ ≤35 ◯ 36-49 ◯ ≥50
2. Residence: ◯ urban ◯ rural
3. Region: ◯ Addis Ababa ◯ Southern Nation ◯ Amhara ◯ Oromia ◯ Others region
4. Smoking Habit: ◯ No ◯ Yes
5. Oral contraceptives: ◯ Not used ◯ Used
6. Alcohol consumption: ◯ No ◯ Yes
7. Tumor size: ◯ ≤ 2 cm ◯ 3–5 cm ◯ > 5cm
8. Obesity: ◯ underweight ◯ normal ◯ overweight
9. Histological grade: ◯ I ◯ II ◯ III
10. Family history of breast cancer: ◯ No ◯ With
11. Stages of breast cancer: ◯ I ◯ II ◯ III ◯ IV
12. Treatments: ◯ Chemotherapy ◯ Radiation ◯ Surgery ◯ Hormone Therapy ◯ Combination of the two or more
13. Breast feeding ◯ No ◯ Yes
14. Patient Status: ◯ (Censored =0) ◯ (Recurrent = 1)
15. Menopause status ◯ Pre-menopausal ◯ Post-menopausal
16. Marital Status: ◯ Married ◯ Single ◯ Widowed ◯ Divorced