FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 2, April, 2023, pp 199 - 205
DOI: https://doi.org/10.33003/fjs-2023-0702-1719

SURVIVAL ANALYSIS OF BREAST CANCER PATIENTS IN KANO STATE

***Aminu Aliyu Nabegu, Dr. Musa Uba Muhammad and Nazir Abdullahi**

Department of Statistics, Faculty of Computing and Mathematical Sciences, Aliko Dangote University of Science and Technology, Wudil.

*Corresponding authors' email: nabegu1001@gmail.com

ABSTRACT

Cancer is a deadly malignant disease and is prevalent in Sub Saharan Africa. Specifically, breast cancer is now the most popular cancer and the second leading cause of death in Nigeria among women. This research is aimed at assessing the length of life of patients and prognostic factors associated with survival of breast cancer patients. Research subjects were breast cancer patients who went to Aminu Kano Teaching hospital (AKTH), patient's follow-up data were obtained from the medical records and pathological variables were obtained from the department of Pathology from 2018- 2022. The survival analysis was performed using Kaplan Meier, log rank and Cox proportional regression model(semi-parametric). The Kaplan Meier results reveals the overall survival to be 0.75(75%). The median survival time is approximately 44 months, that is the probability that half of the patients have died. From the log-rank test, the survival times significantly differ across groups of Tumor stages, Age groups, Lymph node stages and Treatment Types. Results for the Cox proportional hazards model shows that, Treatment type and Tumor stage breast cancer were the risk factor for death in breast cancer patients. It was found that the hazard of death is twice or more for patients with a tumor characterized as tumor stage IV compared to other tumor stages. The hazard of death for patients on Radiotherapy treatment was 1.4 times as patients on Chemotherapy or Hormonal-therapy treatment. The Proportional Hazard assumption was assessed and all the variables meets the assumption.

Keywords: Cox Regression, Time to event Analysis, Survival Analysis, Breast Cancer, Log-Rank Test

INTRODUCTION

Breast cancer is the most commonly diagnosed cancer in women and the most common cause of cancer death in women worldwide. Globally, it is estimated that in 2012 there were 1.68 million new diagnoses (25% of all new cancer diagnoses in women) and 0.52 million deaths (15% of all cancer deaths in women) from invasive breast cancer, corresponding to age-standardized incidence and mortality rates of 43.3 and 12.9 per 100 000, respectively (Ferlay et al., 2013, 2014a). Despite early detection resulting in favorable prognosis, breast cancer is still the leading cause of cancer death among women, especially in economically deprived regions (GLOBOCAN, 2012).

Breast cancer incidence and mortality rates have continuously climbed in some high-income nations while declining in others. (Grybach et al., 2018). However, the incidence and mortality rates of breast cancer have been steadily rising in low- and middle-income nations. (Leal et al., 2016).

Nigeria is currently breast cancer, which is also the most common cancer in the country. It emerged as the condition that required the most thought out of all comparable illnesses, and 56% of respondents named it as one of the top conditions they dreaded the most. Only 32% of participants in a survey on cancer awareness in Nigeria understood that breast lumps are warning indications of cancer, 58.3% were uninformed of the majority of warning signals, 9.8% knew how to detect cancer, and 50% knew that cancer is treatable when identified early. It may be due to the limited awareness of warning indicators and dentification that up to 64% of patients present six months after the onset of symptoms. According to reports, the sickness strikes Nigerian women at a young age. (Ilker et al., 2016). Breast and uterine cancers are the most common types of cancer in Nigeria and its neighboring nations, whereas liver and prostate cancers are more common in men. (Odusanya et al., 2001).

(Aako et al., 2022) conducted a research that concentrated on the variables affecting breast cancer survival rates, Results reveled that the median survival duration of patients was 1,423 days, 72 individuals with breast cancer survived, according to the Kaplan Meier estimator. Survival plot demonstrates that the survival time increases as the likelihood of surviving falls.

In Kano state of Nigeria, the pattern of cancer recorded in its cancer registry for a period of ten years noted a progressive increase in number of cancers cases (Mohammed et al., 2008). This increase is in agreement with the prediction of WHO that there would be a major increase in cancer incidence and mortality in developing countries (WHO, 2005).

Given the aforementioned information, there aren't many studies of survival in areas with poor economies. Therefore, it is appropriate to conduct a thorough study on the Survival analysis of breast cancer in Kano. The goal of this study was to compare the survival of breast cancer patients across levels of potential risk factors and to fit the data into a cox regression model.

MATERIALS AND METHODS

Collet (2003) defines survival analysis as a set of statistical methods for data analysis where the outcome variable is the time until the occurrence of an event of interest. The event of interest can be for example death, occurrence of a disease, failure of a device or recovery from a surgery. General methods of statistical inference and in particular those of regression analysis are not applicable in survival analysis since the analysis is complicated by censoring, that is when complete information on a subject's event is not available. This is the case when the event has not yet occurred at the study termination, or the individual dropped out or observed the event before study termination for reasons unrelated to the study, or the individual was lost to follow- up (Klein and Moeschberger, 2003).

The Survival Function

Let, T be a continuous random variable representing the time until the occurrence of an event of interest, and let f(t) be the

probability density function (pdf) of T. Then its cumulative distribution function is given by

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

(Collet, 2003). Differentiating both sides of (1.1) with respect to t yields the following relationship between survival function and probability density function

$$\frac{dSt}{dt} = -f(t) \quad (2)$$

Two important properties of the survival function derived from (1) are

i. S(t) is a decreasing function on [0, ∞) since F (t) is an increasing function on [0, ∞).
ii. S (0) = 1 and S (∞) = 0 since F (0) = 0 and F (∞) = 1

The Hazard Function

The hazard function or the risk function or intensity rate denoted by h(t) is defined as the rate at which an individual is subject to the event in a small interval of time Δt given that he/she has not observed the event up to time t (Macdonald, 1996). That is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} \quad (3)$$

Note that the hazard function (3) is not a probability, and can theoretically take any value from zero to infinity. By definition of the conditional probability, equation (3) gives which is the relationship between the hazard, survival and probability density functions (Collet, 2003).

Kaplan Meier Estimates (K-M)

The survival probability can be estimated nonparametrically from observed survival times, both censored and uncensored, using the KM (or product-limit) method (Kaplan and Meier, 1958). The idea of this method is based on the probability of the surviving in k or more periods in the study and is a product of k probabilities when each period is observed under it. It is written as

$S(k) = p1 \times p2 \times p3 \times \ldots \times pk$. (Bewick et al., 2004) (4)

As events are assumed to occur independently of one another, the probabilities of surviving from one interval to the next may be multiplied together to give the cumulative survival probability. More formally, the probability of being alive at time (t_j), $S(t_j)$ is calculated from $S(t_{j-1})$, the probability of being alive at (t_{j-1}), (n_j) the number of patients alive just before (t_j), and (d_j) the number of events at (t_j) by

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right) \quad (5)$$

The value of S(t) is constant between times of events, and therefore the estimated probability is a step function that changes value only at the time of each event. This estimator allows each patient to contribute information to the calculations for as long as they are known to be event-free. Were every individual to experience the event (i.e. no censoring), this estimator would simply reduce to the ratio of the number of individuals events free at time t divided by the number of people who entered the study.

The Log Rank Test

Survival in two or more groups of patients can be compared using a nonparametric test. The log rank test (Peto et al., 1977) is the most widely used method of comparing two or more survival curves. The groups may be treatment arms or prognostic groups (e.g. FIGO stage). The method calculates at each event time, for each group, the number of events one would expect since the previous event if there were no difference between the groups. These values are then summed over all event times to give the total expected number of events in each group, say Ei for group i. The log rank test compares observed number of events, say Oi for treatment group i, to the expected number by calculating the test statistic

$$\chi^2 = \frac{\sum_{i=1}^{g}(O_i - E_i)^2}{E_i} \quad \text{(Bewick et al., 2004)} \quad (6)$$

This value is compared to a χ^2 distribution with (g-1) degrees of freedom, where g is the number of groups. In this manner, a P-value may be computed to calculate the statistical significance of the differences between the complete survival curves. If the groups are naturally ordered, a more appropriate test is to consider the possibility that there is a trend in survival across them, for example, age groups or stages of cancer.

The Cox (Semi-Parametric) Proportional Hazards Model

The Cox (proportional hazards or PH) model (Cox, 1972) is the most commonly used multivariate approach for analyzing survival time data in medical research. It is a survival analysis regression model, which describes the relation between the event incidence, as expressed by the hazard function and a set of covariates. A fuller explanation of the hazard function was given in the previous article (Clark et al, 2003). Put briefly, the hazard is the instantaneous event probability at a given time, or the probability that an individual under observation experiences the event in a period centered around that point in time. Mathematically, the Cox model is written as

$$h(t) = h_0(t)e^{(b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_px_p)} \quad \text{,(Fox John, 2002)} \quad (7)$$

where the hazard function h(t) is dependent on (or determined by) a set of p covariates (x_1, x_2, … x_p), whose impact is measured by the size of the respective coefficients (b_1, b_2, …b_p). The term h_0 is called the baseline hazard, and is the value of the hazard if all the x_i are equal to zero (the quantity exp (0) equals 1). The 't' in h(t) reminds us that the hazard may (and probably will) vary over time. An appealing feature of the Cox model is that the baseline hazard function is estimated nonparametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution.

Study Design

The data was a retrospective study of 745 patients collected from medical records and pathology departments respectively of Aminu Kano Teaching Hospital (AKTH), located in Kano State in northwest Nigeria was obtained. The record showed patients that were diagnosed and died of one type of cancer or the other from 2018 to 2022. The nonparametric survival approach was used to estimate the survival probabilities and survival curves, to appraise differences among survival between each of the categorical variables, the log-rank test was applied and checked whether any factor would influence the time to event(death), the semi-parametric regression models, of which the Cox proportional regression model is the most known, provide the relationship of the hazard function to predictors, to validate the use of the cox regression model, a test was also carried out to ascertain the validity of the Proportional hazard assumption using both analytical and graphical methods. All the cancer cases included in the present study were grouped into Tumor Stages, Age of Patients, Tumor Grade, Estrogen Status, Progesterone Status, Treatment type and Lymph Nodes Stage. The data was analyzed using R Statistical Package.

RESULTS AND DISCUSSION

Kaplan-Meier results

From the result in table 1 below, the maximum and minimum survival times for cancer patients were 2 and 52 months respectively. Fifty-two months is the predetermined period for which patients who exceed 52 months after diagnosis are termed to have survived. Figure 1 depict the Kaplan-Meier probability of the survival of the breast cancer patients with a 95% percent confidence bound where it clearly shows that death was higher in the beginning of the follow-up months and it strictly declined in the later months of follow-up. At time zero, the survival probability is 1.0 ($S(0) = 1$),(100% of the patients were alive).After the first year (12 months), the probability of survival was approximately 0.92(92%),the second year (24 months) the probability of survival reduces to 0.77(77%) and towards the end of the follow up period that is 52 months(>4 years) we found out that the probability shrinks to 0.75(75%). This clearly shows us that the as time increases the probability of have survival decreases ($S(\infty) = 0$). The median survival time is approximately 44 months, that is the probability that half of the patients have experienced the event, $s(t) = 0.5$ The K-M plot also revealed that about 70% of the patients experienced the event (death) within the 30 months that is the 75th percentile.

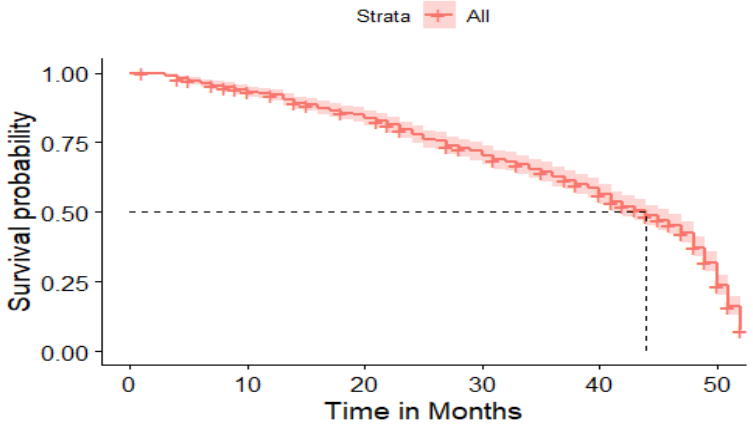


Figure 1: Kaplan-Meier estimates and 95% confidence limits of the survival function for the AKTH Breast Cancer data.

Log-Rank Test

Figure 2 (a) shows that patients with Tumor that has spread to other areas, that is Tumor Stage 4 have shorter survival time with a median survival time of about 33 months, patients who have a small tumor spread(Stage 1) have higher survival time with a median survival time of about 47 months. Again, the plot clearly shows that patients survival decreases with time across all stages of the tumor. With the log-rank test, Table 1 indicates that there is a strong significant difference between Tumor Stage groups. ($\chi_3^2 = 16.2, P = 0.001$).

Figure 2 (b) indicates that the survival outcome is much better for the patients undergoing Hormonal and Chemotherapy treatments with median survival time of 48 and 50 months respectively, the median survival of patients receiving Radiotherapy was about 41 months. Table 1 indicates strong evidence showing statistically significant between the type of treatments received by the patients. ($\chi_3^2 = 9.8, P = 0.007$).

Figure 2 (c) indicates that the survival outcome is slightly better for the patients with tumor grades 1 and 2 respectively compared with patients with a tumor of grade 3. Table 1 indicates that there is no evidence to reject the null hypothesis of equal survival times across the three tumor grade groups. ($\chi_3^2 = 2.7, P = 0.3$).

Figure 2 (d) indicates that the survival outcome is better for lymph node (N_1) patients up to about 42 months, having a median survival time of about 48 months, patients with a lymph node of stage 3(N3) had a poor survival experiences, where they only had a median survival time of less than 40 months. From table 1, gives the log-rank statistics and associated p-values for the three variables of interest is ($\chi_3^2 = 17.4, P = 0.0002$) suggesting that there is no evidence to accept the null hypothesis of no difference in survival time for the three types of lymph Node stages.

Figure 2 (d), indicates a better survival for patients with absence of female hormone Estrogen (negative), however, the patients with the presence of the estrogen hormone also indicates a similar survival level with the negative estrogen for up to about 43 months, after which the survival probability began to drop up to the end of the study period. The Log rank test conducted for Estrogen Status showed no significant difference between the various survival experiences among the categories($\chi_3^2 = 3.9, P = 0.05$) , thus, we failed to reject the null hypothesis.

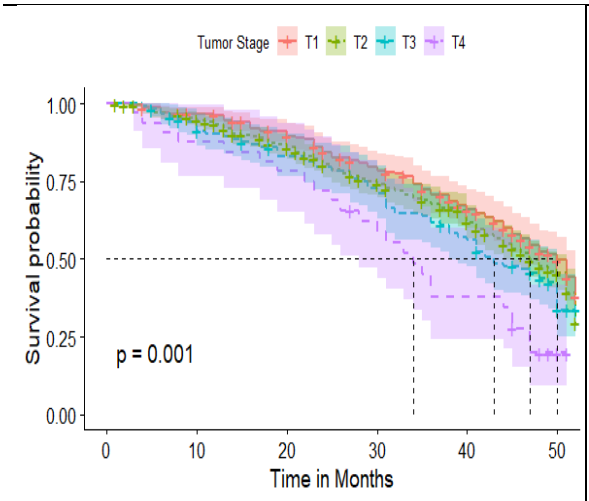
Similarly, Figure 2 (e) shows no substantial differences among the two groups of patients with the presence of progesterone hormones. We can clearly observe that both lines cross each other throughout the study time. The median survival time for patients with positive and negative progesterone hormones are 41 months and 42 months respectively, indicating almost equal survival experiences for all patients. With the log-rank test, Table 1 indicates that there is a no significant difference between the groups. ($\chi_3^2 = 0.5, P = 0.5$), thus, we accept the null hypothesis.

Figure 2 (f) indicates that the survival outcome is better for the patients with a lymph node (stage 1) with a median survival of 46 months, followed by patients in stage 2 with also a median survival time of 46 months, while patients with lymph node of stage 3 had the lowest survival outcome and a median survival time of 36 months, this clearly shows that the higher the lymph node stage the more likely the risk of death due to cancer. Table 1 indicates a clear evidence, that there exists a significant difference across the three lymph node stages ($\chi_3^2 = 17.4, P = 0.0002$).

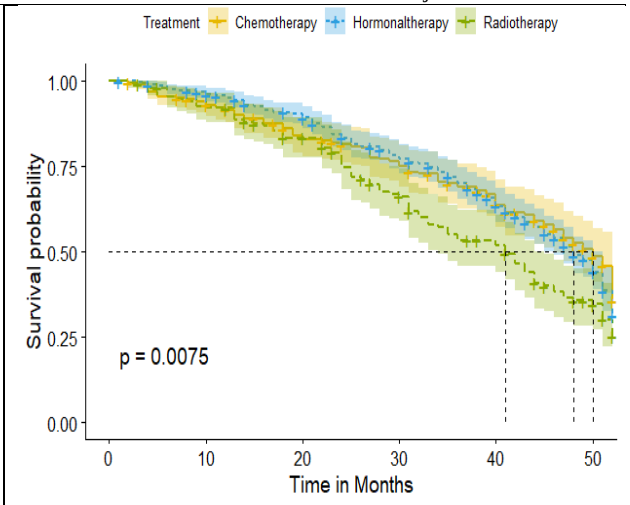
Figure 2 (g) indicates that patients grouped in the age interval >52 years old, have shorter survival time, with a substantial difference among patients grouped in the age interval ≤ 52 years old, with a median survival time of about 48 months, i.e. approximately 3 years. From the log-rank test ($\chi_3^2 = 31.7, P = 0.00000002$), which indicates a clear and strong significance difference between the two age groups.

Table 1: Log-rank test statistics for difference in Survival

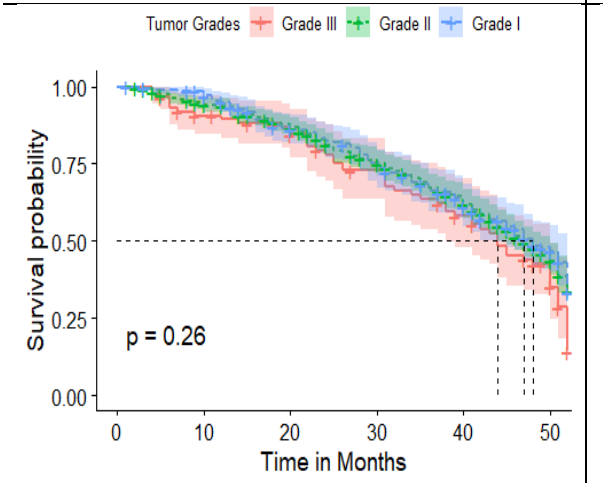
S/n	Variable	Log-rank χ^2 test statistic	(p-value)	Decision
1	Tumor Stages	16.2	0.001	Reject H0
2	Age	31.7	0.00000002	Reject H0
3	Tumor Grade	2.7	0.3	Fail to reject H0
4	Lymph Nodes Stage	17.4	0.0002	Reject H0
5	Estrogen Status	3.9	0.05	Fail to reject H0
6	Progesterone Status	0.5	0.5	Fail to reject H0
7	Treatment	9.8	0.007	Reject H0



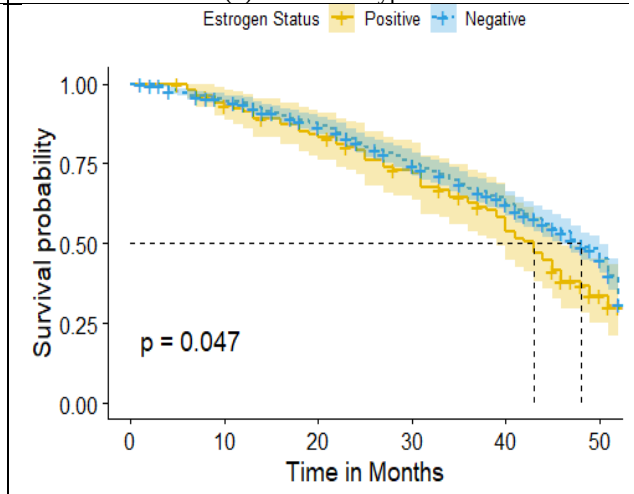
(a) Tumor Stage



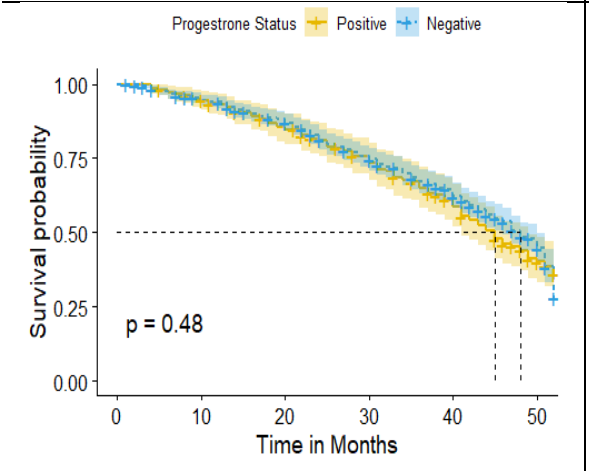
(b) Treatment Type



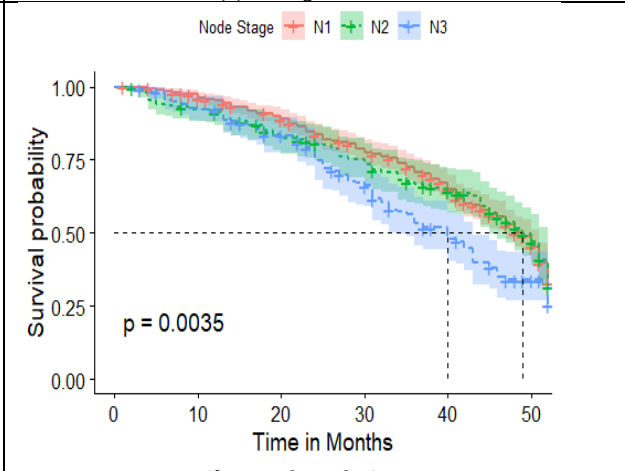
(c) Tumor Grade



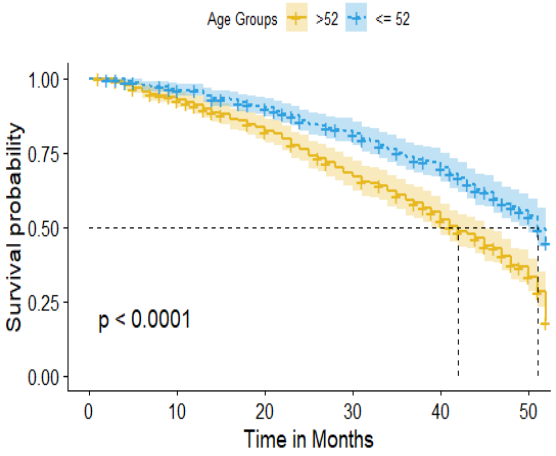
(d) Estrogene Status



(e) Progestrone Status



(f) Lymph Node Status



(g) Age Groups

Figure 2: Plots of the Kaplan-Meier estimates of the survival function for variables under study

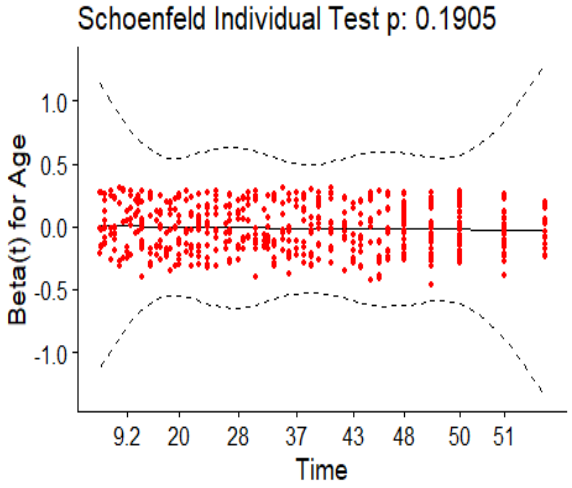
Assessment of the Proportional Hazards Assumption

Figure 3 gives the scaled Schoenfeld residuals versus log-times and the corresponding smooth line on the time scale. The smoothed line in the plot for Age (Figures 3 (a)), appears to have a slope approximately equal to zero. Similarly, for variable Progesterone Status (Figure 3 (b)), Estrogen Status (Figure 3 (c)), Tumor Stage (Figure 3 (d)), Treatment Type (Figure 3 (e)) and Tumor Grade (Figure 3 (f)), and Lymph Node Stage (Figure 3 (g)). This suggests that there may be no time-varying effect of all the variables and this is in agreement with the test. The smoothed line in the plot for the variable gender shows a slight negative slope but the departure from zero slope is not substantial. Therefore, the Schoenfeld residuals indicate that the assumption of proportional hazards is not significantly violated for all the predictor variables retained in the adjusted Cox proportional hazards model.

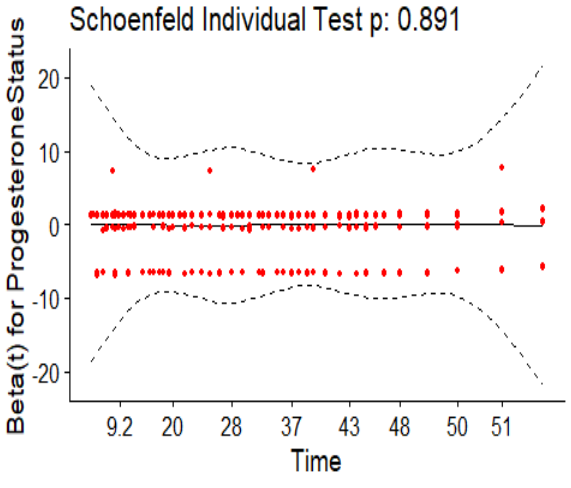
Table 2: Scaled Schoenfeld Residuals of Significant Covariates on the PH

Variables	Chisq	df	p
Age	1.7141	1	0.19
T Stage	1.0031	3	0.80
N Stage	1.9760	2	0.37
Grade	0.7979	2	0.67
Treatment	3.5907	3	0.31
Estrogen Status	0.2403	1	0.62
Progesterone Status	0.0188	1	0.89
GLOBAL	**9.0553**	**14**	**0.83**

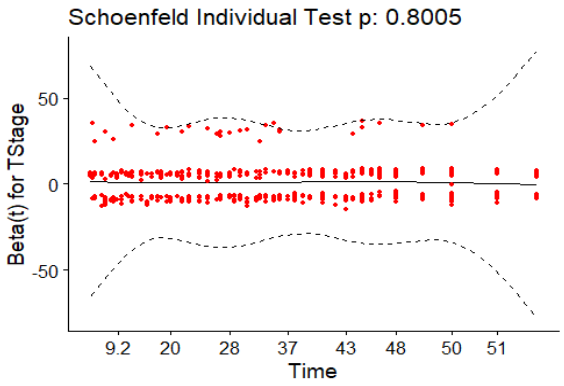
According to these p-values from Table 2, it can be observed that all the p-values for the covariates and the global test are all greater than 0.05. This suggest that the proportional hazards assumption has not been violated by the variables and even from the global test, the p-value of 0.83 shows that a combination of all the variables in the model do not violate the PH assumptions.



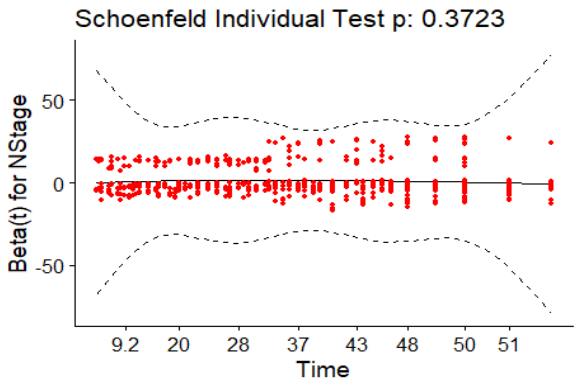
(a) Scaled Schoenfeld residuals plot for Age



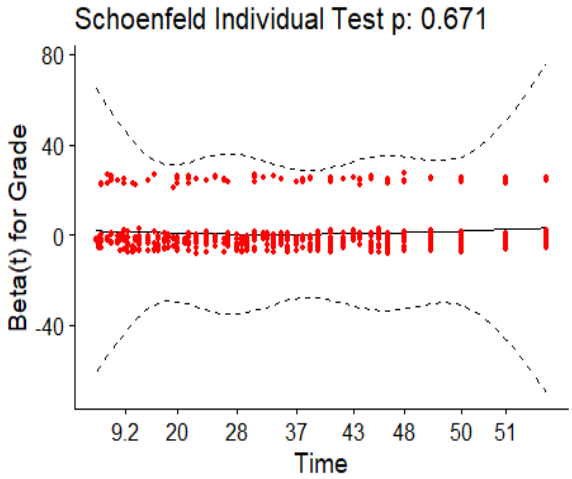
(b) Scaled Schoenfeld residuals plot for Progesterone



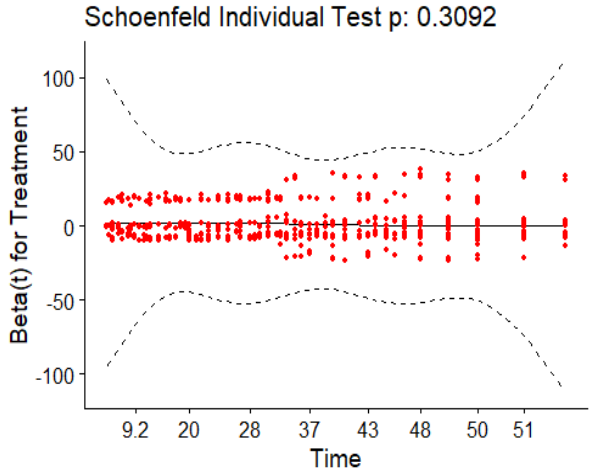
(c) Scaled Schoenfeld residuals plot for Tumor Stage



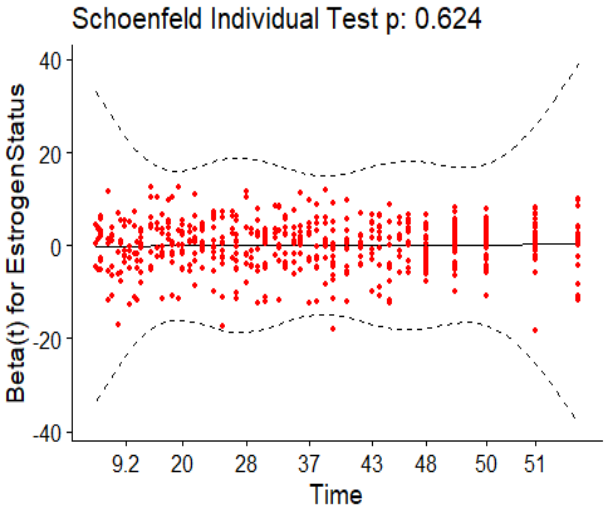
(d) Scled Schoenfield Plot for Node Stage



(e) Scaled Schoenfeld residuals plot for Tumor Grade



(f) Scaled Schoenfeld residuals plot for Treatment type



(g) Scaled Schoenfeld residuals plot for Estrogen Status

Figure 3: Schoenfeld residuals for each explanatory variable versus transformed time for Breast cancer data.

Cox regression

From Table 3, the variables that are not statistically significant at 5% significance level are: Age, Tumor Grades of the breast cancer, Lymph Node, Progesterone Status, Estrogen Status. Treatment types and Tumor Stages of breast cancer were the covariates which were statistically significant at 5% significant level.

Table 3 presents the estimates of the hazard ratios using the Cox proportional hazard model defined in (1.7)**.** The results of the cox model suggest that the hazard of death for patients with Stage 4 Tumor is 2.03 (95% CI: 0.23-3.06) times that of patients with Stage I Tumor, stage II Tumor and Stage III Tumor. The hazard of death for patients undergoing Radiotherapy is 1.4(95% CI: 95% 0.68-2.0) times that of patients undergoing Chemotherapy or Hormonal-therapy.

Table 3: Cox proportional hazards model for all covariates in AKTH data.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Age	-0.007141	0.992884	0.008612	-0.829	0.40699
TstageT2	0.155478	1.168217	0.121869	1.276	0.20203
TstageT3	0.300834	1.350985	0.158683	1.896	0.05798 .
TstageT4	0.711589	2.037226	0.232037	3.067	0.00216 **
NstageN2	-0.002210	0.997793	0.145042	-0.015	0.98785
NstageN3	0.198098	1.219082	0.143291	1.382	0.16682
GradeGrade III	-0.264508	0.767584	0.156362	-1.692	0.09071.
GradeWell differentiated; Grade I	-0.298419	0.741991	0.173802	-1.717	0.08598 .
TreatmentHormonal-Therapy	0.116012	1.123009	0.139446	0.832	0.40544
TreatmentRadiotherapy	0.338799	1.403262	0.168488	2.011	0.04434 *
EstrogenestatusPositive	-0.183519	0.832336	0.250902	-0.731	0.46451
progestronestatusPositive	0.074477	1.077321	0.149667	0.498	0.61875

CONCLUSION

The Kaplan Meier results reveals that, the overall probability of survival was found to be approximately 0.75(75%) with a median survival of 44 months. survival times significantly differ across groups of Tumor stages, Age groups, Lymph node stages and Treatment Types but no significant difference was observed across Tumor grade, Estrogen status and progesterone status. Treatment type and Tumor stage breast cancer were the risk factor for death in breast cancer patients. The results for the Cox proportional hazards model,. It was found that the hazard of death is twice or more for patients with a tumor characterized as tumor stage IV compared to those on either Tumor stage III, II and I. The hazard of death for patients on Radiotherapy treatment was 1.4 times as patients on Chemotherapy or Hormonal-therapy treatment. Probably, patients on the chemotherapy were much healthier than those on the other two treatment types. The Proportional Hazard assumption was assessed and all the variables meets the assumption.

REFERENCES

Aako, O. L., Adewara, J. A and Are, S. O. (2022). Risk Factor Analysis of Breast Cancer Patients in a Nigerian Tertiary Hospital. *FUDMA Journal of Sciences (FJS)*, **6**(3), 95-99. Doi: http://org/10.33003/f js-2022-0603-975

Bewick, V., Cheek, L., Ball, J. (2004). Statistics review 12: survival analysis. Critical care, 8 (5), 389. Cancer. Available from: https://www.ncbi.nlm.nih.gov/books/NBK546553/

Collet, D. (2003). Modeling survival data in medical research, 2nd edition. Chapman & Hall, London.

Cox, D. R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B, 34 (2), 187-220. DOI: 10.15406/bbij.2016.04.00086

Etikan I, Alkassım R, Abubakar S. Statistical analysis on the reported cases of breast cancer. *Biom Biostat Int J.* 2016;4(1):24–26.

Fox, J. (2002). Cox proportional-hazards regression for survival data.

Grybach SM, Polishchuk LZ, Chekhun VF. Analysis of the survival of patients with breast cancer depending on age, molecular subtype of tumor and metabolic syndrome. Exp Oncol. 2018; 40: 243– 248.

IARC Working Group on the Evaluation of Cancer-Preventive Interventions. Breast cancer screening. Lyon (FR): International Agency for Research on Cancer; 2016. 1. Breast

Klein, J. P., and Moeschberger, M. L. (2003). Techniques for censored and truncated data, 2nd edition. Springer-Verlag, New York.

Leal YA, Ferna´ndez-Garrote LM, Mohar-Betancourt A, Meneses-Garcıa A. The importance of registries in cancer control. Salud Publica Mex. 2016; 58: 309–316. https://doi.org/10.21149/spm.v58i2.7802 PMID: 27557391

Macdonald, A. S. (1996). Competing risks, nonparametric and regression models. B A J, 2, 429-448.

Mohammed A., Edino S., Ochicha O., Gwarzo A., Samaila A. A. - *Cancer in Nigeria: a 10-year analysis of the Kano cancer registry*. Niger J Med. 2008; 17: 280-284.

Odusanya O, Olumuyiwa OT. Breast Cancer Knowledge Attitudes and Practice among nurses in Lagos, Nigeria. *Acta Oncol.* 2001;40(7):844– 848.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 35: 1 – 39

World Health Organization - *Global action against cancer*. World Health Organization. 2005.