IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Identification Of Prognostic Factors For Breast Cancer Data Using Survival Analysis Techniques

[1]Srinivasulu V, [2]SC Thasleema, [3]Bharkavi Priya K, [4]Venkataramanaiah M, [5*]**Ahammad Basha Shaik**

[1]Research Scholar, [4]Rtd. Professor, Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.

[3]Statistician, Sri Venkateswara Institute of Medical Sciences, Tirupati, Andhra Pradesh, India

[2]Guest Faculty, Department of Statistics, Vikrama Simhapuri University, Nellore, Andhra Pradesh, India.

[5*]Assistant Professor of Statistics, Department of Community Medicine, Narayana Medical College, Nellore, Andhra Pradesh, *India.*

Abstract: Statistical methods for survival data analysis have continued to flourish in the last two decades. Applications of the methods have broadened from their historical use in cancer and reliability research used in the fields of business, criminology, epidemiology, and social and behavioral sciences. Survival analysis consists of studies of the survival time of a subject (days, weeks, months, or years), which is the time that elapses between the baseline and the moment an adverse event occurs, or the subject drops out of the trial. A retrospective study of 468 breast cancer patterns over a period of one year, i.e., from December, 2021 to November 2022, was collected from the medical records at the Jakka Sujathamma Cancer Detection Centre, Nellore, Andhra Pradesh, India. The log-rank test revealed a statistically significant difference in patient survival times, and it can be deduced that tumor size, family history, advanced stage, and age all have an impact on the disease. The three main risk factors for breast cancer death are stage, grade, and tumor size. Age is one of the most significant risk factors in the Cox PH model analysis; as age rises, so does the likelihood of developing breast cancer.

Keywords: Breast Cancer, Death, Survival analysis, Kaplan-Meier Method, Cox Proportional Hazard model.

INTRODUCTION:

In many clinical studies, the main response variable is often the time to the occurrence of a particular event. In a case of cancer study, for example, surgery, radiation and chemotherapy might be compared with respect to the time from randomization and the start of therapy until death. In this case the event of interest is the death of a patient, but in other situations it might be the end of a period spent in remission from a disease, relief from symptoms, or the recurrence of a particular condition. Such data are generally referred to by the generic term survival data even when the endpoint or event being studied is not death but something else. Questions of interest for such data involve comparisons of survival times for different treatment groups and the identification of prognostic factors useful for predicting survival times. [1,3,8]

Survival analysis can also be used to measure the time to any defined event. Methods for survival analysis allow analysis of such rates without assuming that they are constant. Survival analysis methods are important in trials where participants are entered over a period and have various lengths of follow-up. These methods permit the comparison of the entire survival experience during the follow-up and may be used for the analysis of time to any dichotomous response variable. [6,9,13]

In survival analysis, Survival time can be defined broadly as the time to the occurrence of a given event. This event can be the development of a disease, response to treatment, relapse, or death. The "survival time" refers to a number of years, months, weeks or days from the beginning of the patient observance till the

occurrence of an observed event (death). Therefore, survival time can be tumor-free time, the time from the start of treatment to response, length of remission, and time to death. Survival data can include survival time, response to a given treatment, and patient characteristics related to response, survival, and disease development. [10,20].

Cancer is a group of diseases that cause cells in the body to change and grow out of control. Most types of cancer cells eventually form a lump or mass called a tumor, and are named after the part of the body where the tumor originates. Cancer prevalence is defined as the number of living people who have ever been diagnosed with cancer. It includes people diagnosed with cancer in the past as well those who were recently diagnosed. One in two men and one in three women will be diagnosed with cancer in their lifetime. [2,3,14]. Starting at diagnosis and until the end of treatment, patients with breast cancer receive most of their healthcare from specialists such as a surgeon, medical or radiation oncologist, etc. Once treatment is finished, most of the medical care for these patients will be provided by their primary care clinician. Breast cancer occurs mostly in older women, and many of them have other health issues that are treated by primary care clinicians. [19]

Most breast cancers begin in the parts of the breast tissue that are made up of glands for milk production, called lobules, and ducts that connect the lobules to the nipple. The remainder of the breast is made up of fatty, connective, and lymphatic tissues. Breast cancer is typically detected either during a screening examination, before symptoms have developed, or after a woman notices a lump. Most masses seen on a mammogram and most breast lumps turn out to be benign; that is, they are not cancerous, do not grow uncontrollably or spread, and are not life-threatening. When cancer is suspected, microscopic analysis of breast tissue is necessary for a definitive diagnosis and to determine the extent of spread (in situ or invasive) and characterize the type of the disease. The tissue for microscopic analysis can be obtained via a needle or surgical biopsy. Selection of the type of biopsy is based on individual patient clinical factors, availability of biopsy devices, and resources [2,14,18].

The common sites of cancer in females are cervix uteri's, breast, mouth, ovary, thyroid, vagina, stomach, lung and other types of cancers. Among them, Breast cancer is the most common of all cancers and is the leading cause of cancer deaths among women worldwide. According to estimated number of new cases, both sexes, all ages, in 2020, the breast cancer (13.5%) has held the position of first rank among the top leading cancers when compared with other types of cancers in in India. [19]

An increase in breast cancer incidence in recent years has resulted in a substantial portion of health care dollars being directed towards research in this area. The aims of such research include early detection through mass screening and recurrence prevention through effective treatments. The focus of the investigation was to determine which prognostic measures were predictive of breast cancer recurrence in female patients. Since all breast cancer patients in this report were female, only age and summary distribution of demographic data in the two populations were to be compared. [11,12]. Some statistical notes on survival probabilities (Kaplan–Meier method) have been used. The higher stage, grade, age and history of benign tumor were, the most important risk factors were correlated to mortality in breast cancer patients [2,16]. In a study, comparison of Survival Rates between Chinese and Thai Patients with Breast Cancer conclude that the both Chinese and Thai women need improvement in prognosis, which could conceivably be attained through increased public education and awareness regarding early detection and compliance to treatment protocols [23].

In the past four decades, survival analysis has become one of the most frequently used methods for analyzing data in disciplines ranging from medicine, epidemiology, and environmental health, to criminology, marketing, and astronomy. In many clinical and epidemiological studies, participants or patients may leave the study before it ends and therefore become lost to follow-up.

Inspected by the work done in this direction, an attempt is made in the present study by present authors to compare the survival pattern of patients of different age and income groups suffering with cancer disease. The main aim of the present study is to study some aspects of survival analysis techniques in clinical studies.

MATERIALS AND METHODS:

A retrospective study of 468 breast cancer patterns over a period of one year i.e., from December, 2021 to November, 2022 was collected from the medical records at the Jakka Sujathamma Cancer Detection Centre, Nellore, Andhra Pradesh, India. At this center, different types of treatments are giving for the cancer patients from urban and rural areas.

In this study, survival pattern of cancer patients was studied and survival estimates were calculated using the Kaplan-Meier method. Log rank test was used to test the equality of the groups over the survival distribution estimates. [7, 13,15,22] Cox Proportional Hazard (PH) model with forward conditional method was used to identify the risk factors.

Cox Proportional Hazard (PH) model: **[4,5,8]**

Prognosis plays an important role in medical studies to predict an outcome of a disease. Information about possible prognostic factors can be obtained either from clinical studies designed mainly to identify them, sometimes called prognostic studies, or for ongoing clinical trials that compare treatments as a subsidiary aspect. A prognostic variable (or independent variable) or factor may be either numerical or non-numerical. Cox Proportional Hazards model is also referred to as a relative risk model when the covariates are time-dependent. Cox model does not depend on the true form of the baseline hazard rate function, so the model can be considered as distribution free.

The Cox proportional hazards model possesses the property that different individuals have hazard functions that are proportional, i.e., [h(t|x1)/h(t|x2)], the ratio of the hazard functions of two individuals with prognostic factors or covariates x1 = (x11, x21,…., xp1)′, and x2 = (x12, x22,…., xp2)′ is a constant. This means that the ration of dying of two individuals is the same no matter how long they survive. [Lawless, 1982].

The hazard function given a set of covariates x = (x1, x2,...,xp)′ can be written as a function of an underlying hazard function and a function, say g(x1, ..., xp), of only the covariates, that is,

$$h(t \mid x1,\ldots., xp) = h_0(t)g(x1, \ldots., xp) \text{ or } h(t \mid x) = h_0(t)g(x) \qquad (1)$$

The above hazard function, $h_0(t)$, represents how the risk changes with time, and g(x) represents the effect of covariates. $h_0(t)$ can be interpreted as the hazard function when all covariates are ignored or when g(x) = 1, and is also called the baseline hazard function. The hazard ratio of two individuals with different covariates x_1 and x_2 is

$$\frac{h(t|\mathbf{x}_1)}{h(t\,|\mathbf{x}_2)} = \frac{h_0(t)g(\mathbf{x}_1)}{h_0(t)g(\mathbf{x}_2)} = \frac{g(\mathbf{x}_1)}{g(\mathbf{x}_2)} \qquad (2)$$

Which t is a constant, independent of time.

The Cox proportional hazard model assumes that g(x) in (1) is an exponential function of the covariates, that is,

$$g(x) = e^{\sum_{j=1}^{p} b_j x_j} = e^{\boldsymbol{b}'\mathbf{x}}$$

and the hazard function is

$$h(t\,|\mathbf{X}) = h_0(t)e^{\sum_{j=1}^{p} b_j x_j} = h_0(t)e^{\boldsymbol{b}'\mathbf{x}} \qquad (3)$$

Where b = (b_1, b_2, …, b_p) denotes the coefficients of covariates.

These coefficients can be estimated from the data observed and indicate the magnitude of the effects of their corresponding covariates.

For example, if there is only one covariate treatment, let $x_1 = 0$ if a person receives placebo and $x_1 = 1$ if a person receives the experimental drug. The hazard ratio of the patient receiving the experimental drug and the one receiving placebo based on (2) and (3) is

$$\frac{h(t\,|x_1 = 1)}{h(t\,|x_1 = 0)} = e^{b_1}$$

Thus, the two treatments are equally effective if $b_1 = 0$ and the experimental drug introduces lower (higher) risk of survival than placebo if $b_1<0$ ($b_1>0$). It can be shown that (3) is equivalent to

$$S(t\,|\mathbf{X}) = [S_0(t)]^{e^{\sum_{j=1}^{p} b_j x_j}} = [S_0(t)]^{e^{\boldsymbol{b}'\mathbf{x}}} \qquad (4)$$

Thus, the covariates can be incorporated into the survivorship function.

Dividing both sides of (3) by $h_0(t)$ and taking its logarithm, we obtain

$$\log\frac{h_i(t)}{h_0(t)} = b_1x_{1i} + b_2x_{2i} + \cdots + b_px_{pi} = \sum_{j=1}^{p} b_jx_{ji} = \boldsymbol{b}'\mathbf{X}_{\boldsymbol{i}} \qquad (5)$$

Where the x's are covariates for the i^{th} individual. The left side of (5) is a function of hazard ratio (or relative risk) and the right side is a linear function of the covariates and their respective coefficients.

The use of (5) is used to identify important prognostic factors. In other words, to identify from the p covariates, subsets of variables that affect the hazard rate more significantly and consequently, the length of survival of the patient. We are concerned with the regression coefficients b_i; if b_i is zero, the corresponding

covariate is not related to survival. If b_i is not zero, it represents the magnitude e of the effect of xi on hazard when the other covariates are considered simultaneously. [13,17,21]

To estimate the coefficients, b_1, b_2, …., b_p, Cox proposes a partial likelihood function based on a conditional probability of failure, assuming that there are no tied values in the survival times.

The Cox model for survival provides better estimates of the hazard at any point of time than the Kaplan–Meier method.

Statistical analysis: -

Data has been entered into MS-Excel and statistical analysis was done by using IBM SPSS Version 25.0. For qualitative variables, the values were expressed as number and percentages and for quantitative variables, the values were expressed as mean and standard deviation. To estimate and compare the survival time for different groups, Kaplan–Meier method and the Log rank test have been used. Cox Proportional Hazard (PH) model was used to predict an outcome of a disease, and to get information about possible prognostic factors. All the p values having less than 0.05 were considered as statistically significant.

RESULTS AND DISCUSSION:

Out of 468 breast cancer patients, 267 (57.10%) were alive, 201 (42.90%) were dead patients. Patients' Characteristics at diagnosis by the patient status at the end of follow-up were shown in Table-1 and Pathological characteristics of the breast cancer women were shown in Table-2. Table-1 shows that the Patients' Characteristics at diagnosis status at the end of follow-up for breast cancer patient's data. For the age group, 22.89% of deaths occurred in the age group of less than 30 years, 40.30% of deaths occurred in the age group of 30 to 50 years, and, 36.82% of deaths occurred in the age group of more than 50 years. In financial income, 35.32% of deaths occurred in the poor family income, 54.73% of deaths occurred in the middle class of family income, and, 9.95% of deaths occurred in the higher family income.

Table-2 shows that the pathological characteristics of the breast cancer patient's data. For stage, 64.68% of deaths occurred in the early stage, 21.89% of deaths occurred in the middle stage, and, 13.43% of deaths occurred in the advanced stage. For grade, 16.92% of deaths occurred in the low grade, 41.79% of deaths occurred in the moderate grade, and, 41.29% of deaths occurred in the severe grade. For tumor size, 29.35% of deaths occurred in the less than or equal to 2.0 cms, 47.26% of deaths occurred between the tumor size of 2.0 cms to 5.0 cms, and, 23.38% of deaths occurred in the tumor size of more than or equal to 5.0 cms.

Table-3 showed that the estimation of median survival time (Years) which can have obtained using log-rank test results. The overall median survival time of death of breast cancer patients are six years. The result of Kaplan-Meier curves for estimating the median survival time using log rank test was shown that the age group ($P<0.0001$), financial income ($P<0.0001$), stage ($P<0.0001$), grade ($P=0.028$), and tumor size ($P<0.0001$) were statistically significant relation to the occurrence of death.

Table 1: Patients' Characteristics at diagnosis status at the end of follow-up

Variables	No. of patients	Treatment Outcome	
		Alive (%)	Death (%)
Age Groups			
< = 30 Years	91 (19.44%)	45 (16.85%)	46 (22.89%)
30 – 50 Years	203 (43.38%)	122 (45.69%)	81 (40.30%)
>= 50 Years	174 (37.18%)	100 (37.45%)	74 (36.82%)
Family Income			
Poor	174 (37.18%)	103 (38.58%)	71 (35.32%)
Middle	246 (52.56%)	136 (50.94%)	110 (54.73%)
High	48 (10.26%)	28 (10.49%)	20 (9.95%)

Table 2: Pathological characteristics of the breast cancer

Variables	No. of patients	Treatment Outcome	
		Alive	Death
Stage			
Early	332 (70.94%)	202 (75.66%)	130 (64.68%)
Middle	88 (18.80%)	44 (16.48%)	44 (21.89%)
Advanced	48 (10.26%)	21 (7.87%)	27 (13.43%)
Grade			
Low	83 (17.74%)	49 (18.35%)	34 (16.92%)
Moderate	206 (44.02%)	122 (45.69%)	84 (41.79%)
Severe	179 (38.25%)	96 (35.96%)	83 (41.29%)
Tumor Size (Centimeters)			
<= 2.0	148 (31.62%)	89 (33.33%)	59 (29.35%)
2.0 – 5.0	225 (48.08%)	130 (48.69%)	95 (47.26%)
>= 5.0	95 (20.30%)	48 (17.98%)	47 (23.38%)

Table 3: Median survival time and Log-Rank test for breast cancer patients

Variables	Levels	No. of Patients	Median Survival time (Years)	Log-rank value	P value
Age group (Years)	< = 30 Years	91 (19.44%)	9	53.66	< 0.0001
	30 – 50 Years	203 (43.38%)	7		
	>= 50 Years	174 (37.18%)	6		
Financial Income	Poor	174 (37.18%)	6	43.58	< 0.0001
	Middle	246 (52.56%)	5		
	High	48 (10.26%)	10		
Stage	Early	332 (70.94%)	10	45.97	< 0.0001
	Middle	88 (18.80%)	6		
	Advanced	48 (10.26%)	4		
Grade	Low	83 (17.74%)	6	7.16	0.028
	Moderate	206 (44.02%)	4		
	Severe	179 (38.25%)	2		
Tumor size (Cms)	<= 2.0	148 (31.62%)	7	55.40	< 0.0001
	2.0 – 5.0	225 (48.08%)	5		
	>= 5.0	95 (20.30%)	3		

The Kaplan-Meier survival curve for (a) Age group, (b) Stage, (c) Grade, and (d) Tumor size was shown in Figure-1. The results of Cox Proportional Hazard (PH) model analysis shows that hazard ratio for death due to breast cancer in women with a age group (≤30 Years Vs ≥ 50 Years: Hazard Ratio = 3.33, 95% CI; 1.85-5.99), stage (Early Vs Advanced: Hazard Ratio = 12.71, 95% CI; 4.94-32.66), grade (Low Vs Severe: Hazard Ratio = 0.18, 95%CI; 0.09-0.37), family income (Poor Vs High: Hazard Ratio = 2.20, 95%CI; 1.33-3.62) and tumor size (≤ 2.0 cm Vs ≥ 5.0 cm: Hazard Ratio = 6.34, 95%CI;2.67-15.05) were identified as the risk factors for breast cancer patients data.

The authors studied some applications of survival techniques using breast cancer data. Current trends point out that a higher proportion of the disease is occurring at a younger age in Indian women, as compared to the West. The National Cancer Registry Program analyzed data from cancer registries for the period from 1988 to 2013 for changes in the incidence of cancer. All population-based cancer registries have shown a significant increase in the trend of breast cancer. [19]. In this study, the age of patients, stage, size of tumor, and tumor grade are the common risk factors that can be used to predict the survival probability related with both events. These risk factors are consistent with the result in other studies. [14]. The analysis performed is from a statistical point of view; clinical justification needs to be considered in practice.

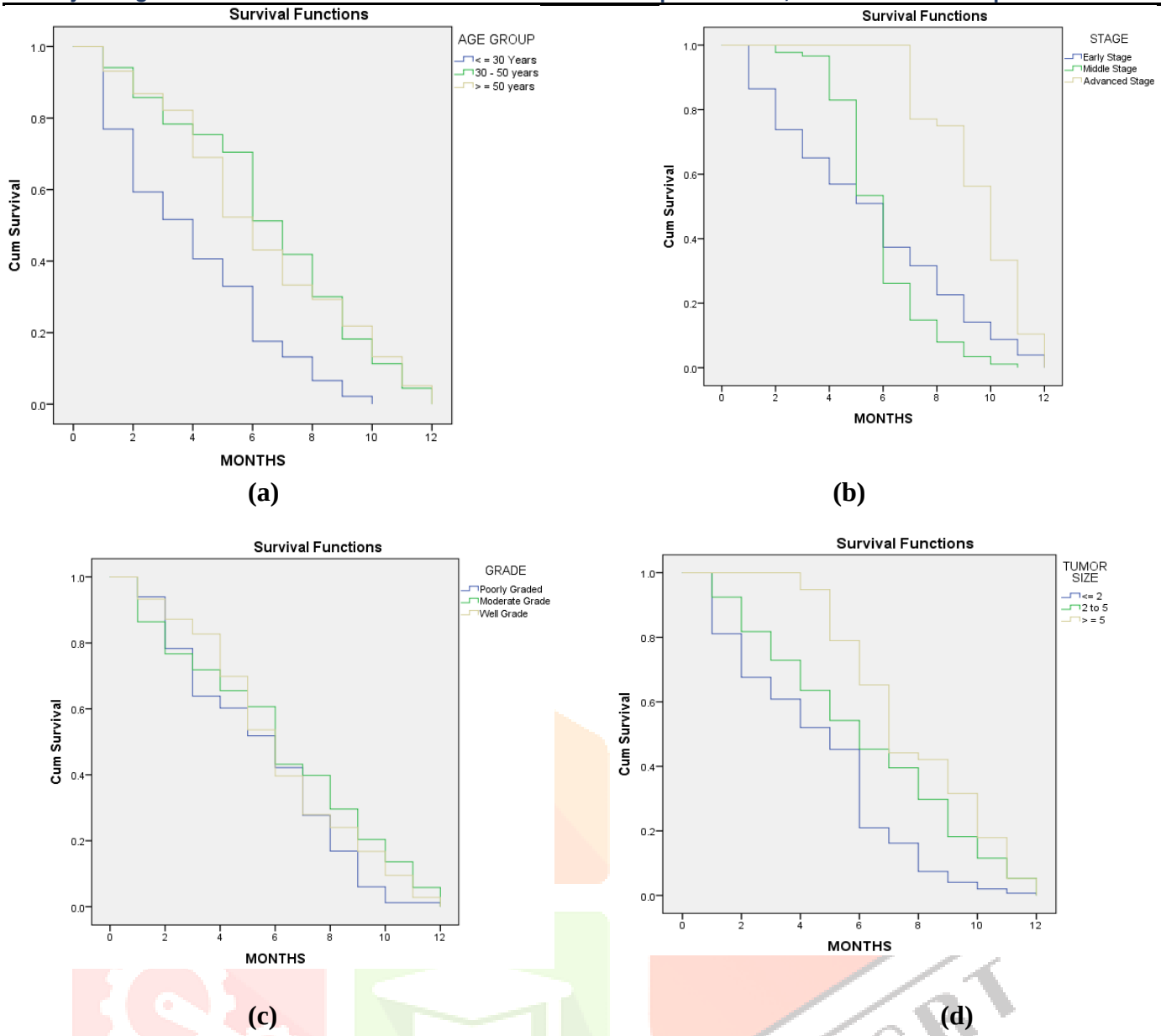

Figure 1: Kaplan-Meier Estimate of Survival Curves for (a) Age Group, (b)Stage, (c) Grade, and (d) Tumor size

Table 4: Identifying Prognostic Factors for breast cancer data using Cox regression model

Variable Levels	Hazard Ratio (HR)	95% C.I. of HR		P-value
		Lower	Higher	
30-50 Years Vs ≥ 50 Years	1.23	0.73	2.12	0.448
≤ 30 Years Vs ≥ 50 Years	3.33	1.85	5.99	< 0.001
Early Vs Advanced	12.71	4.94	32.66	< 0.001
Middle Vs Advanced	2.98	1.56	5.72	0.001
Poor Vs High	2.20	1.33	3.62	0.002
Low Vs Severe	0.18	0.09	0.37	< 0.001
Moderate Vs Severe	0.03	0.013	0.073	< 0.001
≤ 2.0 Vs ≥ 5.0	6.34	2.67	15.05	0.001
2.0-5.0 Vs ≥ 5.0	1.34	0.69	2.60	0.383

CONCLUSIONS:

The present work is dedicated to study some aspects of survival analysis techniques in clinical studies. With the Kaplan–Meier survival analysis procedure, we have examined the distribution of time to effect for two or more different groups. The log-rank test showed that there is a statistically significant difference in survival times of patients and it can be concluded that age, higher stage, grade, family history and tumor size influence the disease. In the analysis of Cox PH model, Age is one of the most important risk factors; as age increases, the risk of being diagnosed with breast cancer also increases. Stage, Grade, and tumor size are the most important risk factors related to death in breast cancer.

REFERENCES:

[1] Altman DG. 1991. Practical Statistics for Medical Research. New York: Chapman and Hall, PP.383-392.

[2] Rajaeefard AR, MR Baneshi, AR Talei, D Mehrabani. 2009. Survival Models in Breast Cancer Patients. Iranian Red Crescent Medical Journal, 11(3):295-300.

[3] Armitage P. 1977. Statistical Methods in Medical Research. New York: John Wiley and Sons.

[4] Cox DR. 1972. Regression models and life tables. J R Stat Soc Series B Stat Methodol, 34:187-202.

[5] Cox DR., Oakes D. 1984. Analysis of Survival Data. Chapman & Hall, New York.

[6] Gross AJ, Clark VA. 1975. Survival Distributions: Reliability Applications in the Biomedical Sciences. Wiley, New York.

[7] Kaplan E, Meier P. 1958. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 53: 457-481.

[8] Lawless, JF. 1982. Statistical Methods and Model for Lifetime Data. Wiley, New York.

[9] Mann NR, Schafer RE, Singpurwalla, ND. 1974. Methods for Statistical Analysis of Reliability and Life Data. Wiley, New York.

[10] Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst, 22:719–748.

[11] Miller RG, Jr. 1981. Survival Analysis. New York: John Wiley and Sons.

[12] Nazera Khalil Dakhil, Yahya Mahdi Al–Decemberali, Muna Abbas Mseer Al–A'bidy Al. 2012. Analysis of Breast Cancer Data using Kaplan–Meier Survival Analysis. Journal of Kufa for Mathematics and Computer,1(6):7-14.

[13] Nelson W.1972. Theory and Applications of Hazard Plotting for Censored Failure Data. Technometrics,14:945-966.

[14] Ahammad Basha Shaik, M. Venkataramanaiah, S.C. Thasleema. 2015. Statistical Applications of Survival data analysis for breast cancer data. i-manager's Journal on Mathematics, 4(2): 30-36.

[15] Nelson W 1982. Applied Life Data Analysis. Wiley, New York.

[16] Rajaeifard AR, Talei AR, Baneshi MR. 2005. Survival analysis models for breast cancer patients in Shiraz 1993-2002. Journal of Medical Research, 3(4):41-50.

[17] Woolson R. 1987. Statistical Methods for the Analysis of Biomedical Data. New York: John Wiley and Sons.

[18] Pavana Sandhya, Shanthi. M, Nusrath Fareed, KM Sudhir, RVS Krishnakumar.2011. Retrospective Analysis of Hospital Records at a Cancer Institute in Nellore District, Andhra Pradesh. Journal of the Indian Association of Public Health Dentistry,18:161-166.

[19] Mehrotra R, Yadav K. 2022. Breast cancer in India: Present scenario and the challenges ahead. World J Clin Oncol, 13(3):209-218

[20] Elisa T Lee and Oscar T Go. 1997. Survival Analysis in Public Health Research, Annu. Rev. Public Health, 18:105–34.

[21] Altman DG. 1991. Practical Statistics for Medical Research. Chapman & Hall, New York.

[22] Bland JM, Altman DG. 1998. Survival probabilities (the Kaplan- Meier method), BMJ, 317: 1572.

[23] Yanhua Che, Jing You, Shaojiang Zhou, Li Li, Yeying Wang, Yue Yang, Xuejun Guo, Sijia Ma, Hutcha Sriplung. 2014. Comparison of Survival Rates between Chinese and Thai Patients with Breast Cancer. Asian Pacific Journal of Cancer Prevention, 15: 6029-6033.