Global Scientific and Academic Research Journal of Multidisciplinary Studies
ISSN: 2583-4088 (Online)
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journals-gsarjebm-home/

UNDERSTANDING THE KAPLAN-MEIER ESTIMATE FOR BREAST CANCER - A RETROSPECTIVE STUDY

BY

Velu Chinnasamy Shanmugam

Principal, Advanced Analytics, DIRECTV, Irving, Texas, United States-75063

Article History

Received: 03/07/2023
Accepted: 14/07/2023
Published: 16/07/2023

Vol – 2 Issue – 7

PP: - 09-13

Abstract

Breast cancer is a cancerous tumor that develops from breast cells. Although it mostly affects women, men can occasionally develop breast cancer. For estimating the median of the distribution of breast cancer patients' survival periods after being enrolled in the study, Kaplan-Meier estimator is more accurate method to be implemented. The factors included in the study: age, race, marital status, patient differentiation, estrogen status, progesterone status, disease stage, and the outcomes of the therapy of 4024 breast cancer patients. The median overall survival time for breast cancer patients was 60 months, with a mean age of 53.97 and a standard deviation of 8.963. According to this, 78% of breast cancer patients survived for more than 60 months after their diagnosis.

Keywords: *Survival Analysis, Kaplan-Meier, Breast Cancer, Stages, Omnibus Tests, and Long-Run Test.*

INTRODUCTION

Breast cancer is a malignant tumor that develops from breast cells. Although it mostly affects women, men can occasionally develop breast cancer[1]. If caught early enough, breast cancer is highly treatable. Age at diagnosis, race, cancer stage at diagnosis, lymph node status, type of treatment, immunohistochemistry subtype, nuclear grade, histological grade, access to care, and environmental factors are a few of the variables that affected the survival of breast cancer patients[6]. Age at diagnosis, ethnicity, cancer stage at diagnosis, lymph node status, therapy used, immunohistochemistry subtype, nuclear grade, histological grade, access to care, and environmental factors are only a few of the variables that affected the survival of breast cancer patients. According to the World Health Organization[2], breast cancer affects 2.1 million women annually and is the most often diagnosed disease in women. Most Asian nations reported an increase in the frequency of breast cancer[2,5,8]. Breast cancer is the most common type of cancer and the one that kills Malaysian women the fastest[4].

A statistical tool for analyzing breast cancer data is the Kaplan-Meier (KM) method, also known as the product limit method[3]. It is used to examine the survival time distribution of patients who have been enrolled in the study. This is expressed in the analysis as the percentage of patients who were still alive at a particular point after being enrolled in the research or recruited for it. The nonparametric maximum likelihood estimator is another name for the KM estimator. It is employed to calculate the likelihood of survival. The approach determines the likelihood of dying depending on whether you survive up until that point in time. In epidemiology, the Kaplan-Meier survival curve is used to compare two groups of people and assess time-to-event data. The survival curve is used to calculate the percentage of patients who will survive a particular event, such as death within a certain time frame. This can be computed for two patient or subject groups, as well as their statistical difference in survival rates.

MATERIALS AND KM METHODS

After lung cancer, breast cancer is one of the most prevalent and feared cancers that cause death. It is a significant contributor to cancer-related morbidity and mortality in women[7]. The 4024 breast cancer cases represented by the study's data were located in the Kaggle database. This database of breast cancer patients was acquired from the SEER Program of the NCI's November 2017 update, which offers details on population-based cancer statistics. The dataset included female patients who had been diagnosed between 2006 and 2010 with infiltrating ductal and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3). In the end, 4024 patients were

*Corresponding Author: **Velu Chinnasamy Shanmugam**

included after patients with uncertain tumor sizes were excluded, positive regional LNs, patients with survival times of less than one month, and more.

Age, sex, occupation, disease stage, hospital stay duration, patient status, and treatment outcomes were the covariates or independent variables taken into account in the study. The K-M approach is the suggested technique for survival analysis in cancer trials. It is a technique for calculating the proportion of patients who survive for a specific amount of time after therapy. It is put into practice by examining the distribution of patient survival times after enrollment in the study. Following their enrollment, the analysis expresses them in terms of the percentage of patients who are still living. Without making any assumptions about the underlying probability distribution, the K-M approach is used to estimate the survival curves for patients from the recorded survival times. The method is founded on the fundamental tenet that the likelihood of surviving P or more periods from the time of entry into the research is the product of the P observed survival rates for each period, or the cumulative surviving, and is denoted by:

$$S(P) = (K_1), (K_2), (K_3), \dots (K_n) \quad \dots (1)$$

K_1 = Proportion of surviving the initial stage,
K_2 = The proportion of individuals who survive past the second stage, and so on.

The proportion of surviving a period i, having survived up to period i is given by,

$$K = \frac{r_i - d_i}{r_i} \quad \dots (2)$$

r_i = Alive numbers at the initial stage
d_i = The number of deaths within the stages

The log-rank test is a statistical hypothesis test that may be used to compare two survival curves. It is employed to investigate the null hypothesis that there is no significant difference in the population survival curves.
The test statistic is calculated by:

$$\chi^2(Long - rank) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots \quad \dots (3)$$

Where O_1 and O_2 are the total number of observed events in groups 1 and 2, respectively, E_1 and E_2 are the total number of expected events in the respective groups.

RESULT AND DISCUSSIONS

Table 1: Summary results of Breast Cancer Censored and death events by different demographic, health, and risk behavior variables.

Variables	Status		Total (%)
	Number of Deaths (%)	Number censored (%)	
Marital Status: Married	358 (13.5)	2285 (86.5)	2643 (65.7)
Single	104 (16.9)	511 (83.1)	615 (15.3)
Separated	15 (33.3)	30 (66.7)	45 (1.1)
Widowed	49 (20.9)	186 (79.1)	235 (5.8)
Divorced	90 (18.5)	396 (81.5)	486 (12.1)
Race: White	510 (14.9)	2903 (85.1)	3413 (84.8)
Black	73 (25.1)	218 (74.9)	291 (7.2)
Others	33 (10.3)	287 (89.7)	320 (8.0)
Stage: Regional	581 (14.8)	3351 (85.2)	3932 (97.7)
Distant	35 (38.0)	57 (62.0)	92 (2.3)
T-Stage: T1	157 (9.8%)	1446 (90.2)	1603 (39.8)
T2	303 (17.0)	1483 (83.0)	1786 (44.4)
T3	116 (21.8)	417 (78.2)	533 (13.2)
T4	40 (39.2)	62 (60.8)	102 (2.5)
N-Stage: N1	270 (9.9)	2462 (90.1)	2732 (67.9)
N2	165 (20.1)	655 (79.9)	820 (20.4)
N3	181 (38.3)	291 (61.7)	472 (11.7)
Six-Stage: IIA	96 (7.4)	1209 (92.6)	1305 (32.4)
II B	135 (11.9)	995 (88.1)	1130 (28.1)

*Corresponding Author: **Velu Chinnasamy Shanmugam**

III A	184 (17.5)	866 (82.5)	1050 (26.1)
III B	20 (29.9)	47 (70.1)	67 (1.7)
III C	181 (38.3)	291 (61.7)	472 (11.7)
Differentiate: Moderately	305 (13.0)	2046 (87.0)	2351 (58.4)
Poorly Differentiated	263 (23.7)	848 (76.3)	1111 (27.6)
Undifferentiated	9 (47.4)	10 (52.6)	19 (0.5)
Well Differentiated	39 (7.2)	504 (92.8)	543 (13.5)
Estrogen Status: Negative	108 (40.1)	161 (59.9)	269 (6.7)
Positive	508 (13.5)	3247 (86.5)	3755 (93.3)
Progesterone Status: Negative	204 (29.2)	494 (70.8)	698 (17.3)
Positive	412 (12.4)	2914 (87.6)	3326 (82.7)
Grade: Grade-1	39 (7.2)	504 (92.8)	543 (13.5)
Grade-2	305 (13.0)	2046 (87.0)	2351 (58.4)
Grade-3	263 (23.7)	848 (76.3)	1111 (27.6)
Grade-4	9 (47.4)	10 (52.6)	19 (0.5)

From Table 1 we observed that, under the marital status the more number of deaths 33.3% happened for separated people and less number of deaths occurred 13.5% for married peoples. Black race people had 25.1% number of deaths with the maximum and other race people with minimum 10.3% number of deaths. The distant stage people were more in the number of deaths 38% compared with regional 14.8%. Under the T-stage, Stage T4 found with the most number of deaths 39.2%, and T1 found the minimal number of deaths 9.8%. From the N-stage group, 38.3% was observed in N3 stage with most number of deaths. In the Six-stage the most number of deaths occurred at IIIC with 38.3%. Undifferentiated people found with the most number of deaths 47.4%. In the estrogen status negative found with 40.1% of deaths compared with positive deaths with 13.5%, same scenario observed under progesterone status with negative deaths 29.2% more than positive deaths 12.4%. The grade 4 (47.4%) people found with more number of deaths compared with other three grades.

Table 2: Summary on demographic variables

Demographic	Mean±SD
Age	53.97 ± 8.963
Tumor Size	30.47 ± 21.120
Regional Node Examined	14.36 ± 8.100
Regional Node Positive	4.16 ± 5.109
Survival Months	71.30 ± 22.921

The survival and hazard curves were estimated and plotted using Kaplan-Meier. The patients with breast cancer were 53.79 years old on average. Tumor size observed 30.47 mean score and 21.120 standard deviation. The average survival month of the patients is 71.30 as observed in table 2.

Table 3: Summary Statistics for Risk Factors Used in Risk Model

Variables	β	SE	Wald	P-Value	Exp (β)	95.0% CI for Exp (β)	
						Lower	Upper
Marital Status: Married	.055	.055	1.006	.316	1.057	.949	1.177
Single	.007	.067	.010	.920	1.007	.882	1.149
Separated	.279	.190	2.141	.143	1.321	.910	1.919
Widowed	.022	.090	.060	.807	1.022	.857	1.220

*Corresponding Author: **Velu Chinnasamy Shanmugam**

Race: White	.031	.062	.241	.624	1.031	.912	1.165
Black	.099	.091	1.179	.277	1.104	.923	1.320
Stage Regional & Distant	-.011	.150	.005	.943	.989	.737	1.328
T-Stage: T1	.268	.278	.926	.336	1.307	.758	2.253
T2	.295	.273	1.167	.280	1.343	.787	2.292
T3	.231	.277	.692	.406	1.260	.731	2.170
N-Stage: N1	.186	.319	.340	.560	1.204	.645	2.250
N2	.026	.312	.007	.935	1.026	.557	1.891
Six-Stage: IIA	-.168	.316	.283	.595	.845	.455	1.571
II B	-.173	.312	.305	.581	.842	.456	1.552
III A	-.005	.307	.000	.986	.995	.545	1.817
Differentiate: Moderately	-.062	.050	1.544	.214	.939	.851	1.037
Poorly Differentiated	-.116	.058	4.018	.045	.890	.795	.997
Well Differentiated	-.526	.321	2.685	.101	.591	.315	1.109
Estrogen Status	-.011	.150	.005	.314	.913	.764	1.090
Progesterone Status	-.091	.091	1.013	.004	.856	.769	.952
Age	.000	.002	.051	.821	1.000	.996	1.004
Regional Node Examined	.002	.002	1.135	.287	1.002	.998	1.007
Regional Node Positive	-.004	.008	.208	.649	.996	.981	1.012

Table-3: displays the final multivariate model. In addition, 12 of the 2 variables had a significant influence on patients with specific metastasis (Poorly Differentiated and Progesterone Status), whereas the remaining 10 were excluded from the model due to non-significant effect on the chance of a conversion.

Table 4: Omnibus Tests of Model Coefficients

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	χ^2	df	Sig.	χ^2	df	Sig.	χ^2	df	Sig.
49221.208	34.990	20	.020	36.186	20	.015	36.186	20	.015

a. Beginning Block Number 1. Method = Enter

The Omnibus Tests of Model Coefficients is used to check that the categorical variables are an improvement over the baseline model. It uses chi-square tests to see if there is a significant difference between the Log-likelihoods of the baseline model and the new model. If the new model has a significantly reduced -2 Log-likelihoods compared to the baseline then it suggests that the new model is explaining more of the variance in the outcome and is an improvement! Here the chi-square is highly significant (Chi-Square = 34.990, df = 20, p<.000) so our new model is significantly better.

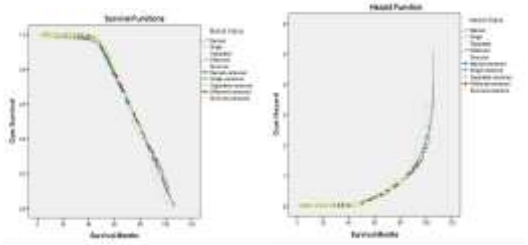

Figure-1: Survival & Hazard functions of Breast Cancer patients according to Marital Status.

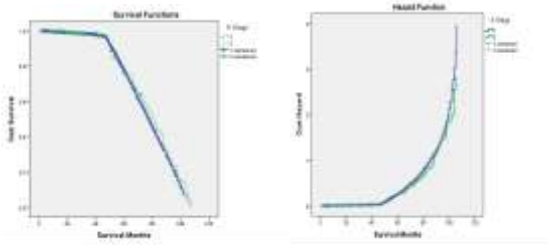

Figure-2: Survival & Hazard functions of Breast Cancer patients according to Stage wise.

*Corresponding Author: **Velu Chinnasamy Shanmugam**

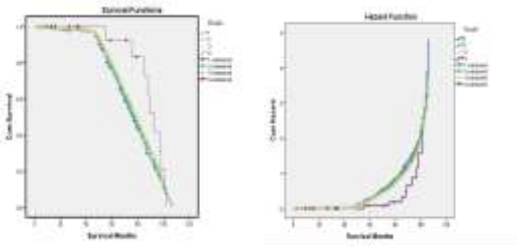

Figure-3: Survival & Hazard functions of Breast Cancer patients according to Grade wise.

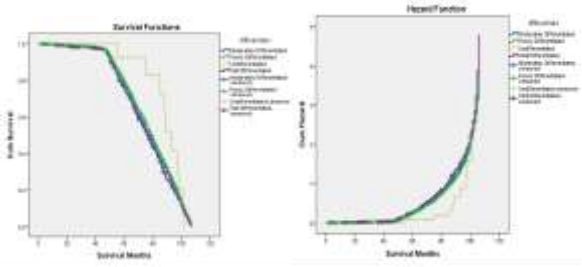

Figure-4: Survival & Hazard functions of Breast Cancer patients according to Differentiated.

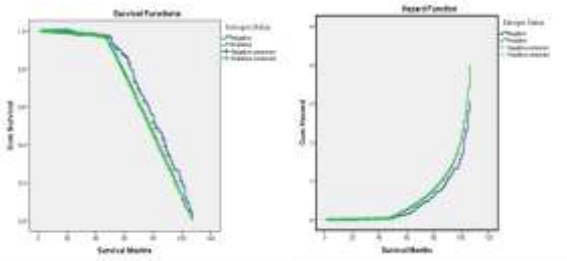

Figure-5: Survival & Hazard functions of Breast Cancer patients according to Estrogen Status.

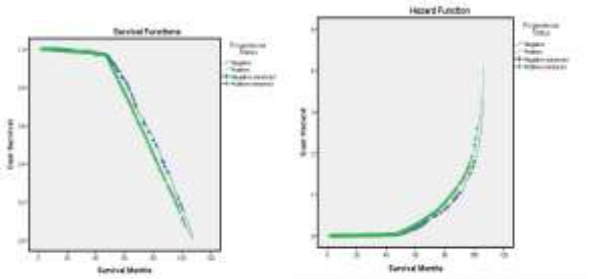

Figure-6: Survival & Hazard functions of BC patients according to Progesterone Status.

CONCLUSIONS

In the study of time-to-event data, the statistical method known as Kaplan-Meier is particularly helpful in the field of epidemiology. The technique is used in survival analysis to examine breast cancer patients who have reached a particular event and those who have been censored over a predetermined time period. It is also highly useful for comparing participant groups, such as the control group and the treatment group. Statistical applications like SPSS, Strata, SAS, R, and Python can be used to analyze data and create useful tables like the overall comparisons table as well as the Kaplan-Meier estimate curve. The KM estimate is also used in fields like engineering, economics, physics, and others.

REFERENCES

1. Breast Cancer, Medicine Net, (2018). Available: https://www.medicinenet.com/breast_cancer_facts_stages/article.
2. Cancer: Breast cancer, World Health Organization, (2018). Available: http://www.who.int/cancer/prevention/diagnosisscreening/breast-cancer/en/.
3. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Amer Statist Assoc. (1958). Vol.53, (282), pp.457-81.
4. N.A.Abdullah et.al, Survival Rate of Breast Cancer Patients in Malaysia: A Population-based Study, Asian Pacific J. Cancer Prev., (2013). Vol. 14 (8), pp. 4591–4594.
5. N.B.Pathy et.al, Ethnic differences in survival after breast cancer in South East Asia, PLoS One, (2012). Vol. 7 (2), pp.1-6.
6. N.Nordin, N.M.Yaacob, N.H.Abdullah, and S.M.Hairon, Survival Time and Prognostic Factors for Breast Cancer among Women in North-East Peninsular Malaysia, Asian Pacific J. Cancer Prev., Vol. 19 (2), pp. 497–502, 2018.
7. Ries LA, Wingo PA, Miller DS, Howe HL, Weir HK, Rosenberg HM, The annual report to the nation on the status of cancer, 1973-1997, with a special section on colorectal cancer. Cancer (2000). Vol.88, pp.2398-2424.
8. S.K.Park, Y.Kim, D.Kang, E.J.Jung, and K.Y.Yoo, Risk factors and control strategies for the rapidly rising rate of breast cancer in Korea, J. Breast Cancer, (2011). Vol. 14 (2), pp.79-87.
9. V.M.Medina, A.Laudico, M.R.Mirasol-Lumague, H.Brenner and M.T.Redaniel, Cumulative incidence trends of selected cancer sites in a Philippine population from 1983 to 2002: A join point analysis, Br. J. Cancer, (2010). Vol. 102 (9), pp. 1411–1414.

*Corresponding Author: **Velu Chinnasamy Shanmugam**