Analysis of Breast Cancer Data using Kaplan–Meier Survival Analysis

Nazera Khalil Dakhil Yahya Mahdi Al–Decemberali Muna Abbas Mseer Al–A'bidy
College of Mathematics and Computer Sciences
University of Kufa

.

Abstract

The Kaplan–Meier estimator is a very popular that provides better estimates to determine the median when the sample size is reasonably large. The aim of this research was mainly concerned with a study and analysis an estimation of the survivorship time of real data of breast cancer patients in Iraq.

Keywords: Survival analysis, censoring, Kaplan–Meier, log–rank test.

Introduction

The Kaplan–Meierprocedure is a method of estimating time–to–event models in the presence of censored cases. It's an intrinsic characteristic of survival data is the possibility for censoring of observations (that is, the actual time until the event is not observed). Such censoring can arise from withdrawal from the experiment or termination of the experiment[8]. The Kaplan–Meiermodel is based on estimating conditional probabilities at each time point when an event occursand taking the product limit of those probabilities to estimate the survival rate at each point in time. The Kaplan–Meierestimator december be obtained as the limiting case of the classical actuarial estimator, and it seems to have been first proposed by (Bohmer 1912)[5]. Kaplan and Meier (1958) were the first who carried out the solution of a problem to estimate the survival curve in a simple way while considering the right censoring.

Bland and Altman (1998)[4]contained some statistical notes on survival probabilities (Kaplan–Meiermethod).*Tovey and et al*(2009)[9] presented Kaplan–Meier survival curves by using breast cancer–specific death as an outcome endpoint (log–rank testing).*Rajaeefard and et al* (2009)[7] they concluded that the higher stage, grade, age and history of benign tumor were, the most important risk factors were correlated to mortality in breast cancer patients.*Zino*(2010)[10] examined a potential role for sirtuins in breast cancer disease (including anti–tumor treatment). The Kaplan–Meier analysis and Cox regression analysis demonstrated the relative

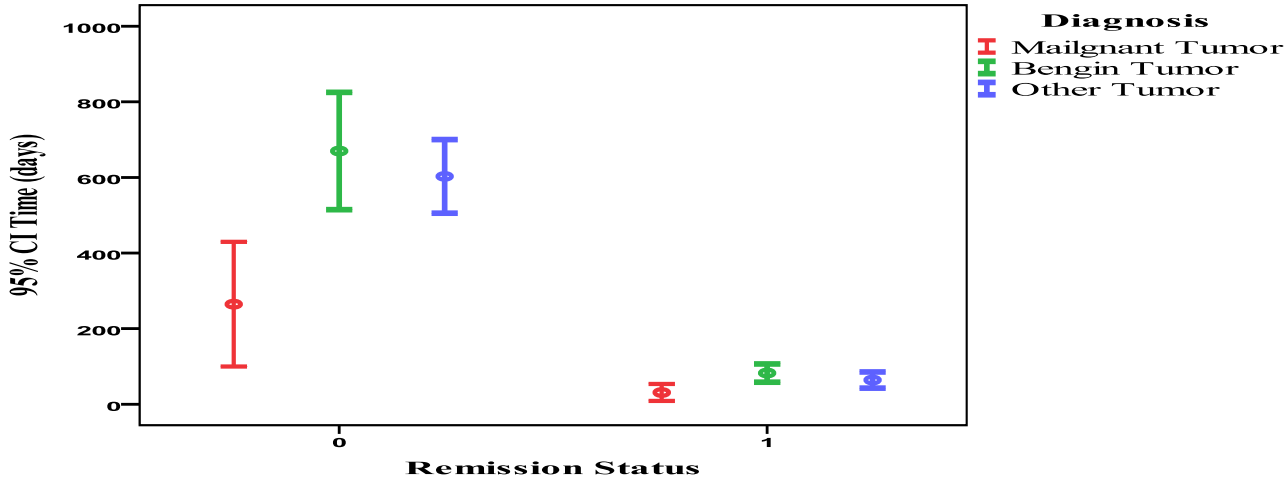
pathological prognostic markers. And to estimate the survivorship function of three distinct groups; malignant, benign, and other tumors for the breast tumor patients by using the Kaplan–Meiermethod and to make comprise by using the log–rank testwas studied by*Al–A'bidy* (2011)[2].

The summarize of this paper is as follows. Section 2 reviews the data and method and section 3 explain the Kaplan–Meiermethod and log–rank test. Section 4 displays analysis of results to compare between breast tumor groups and this is followed by the conclusion in section 5.

1. *Data and Methods*

The collected simple random sample data was the specialized breast diseasesclinic in Al–Sadder Medical City in Al–Najaf. All the cases included in present study were diagnosed as either malignant, benign, and other tumors. Patients were followed up as a minimum to one year. In survival analysis, follow up periods were calculated from the first consultation with surgeon.

The data consisted of 254 women from year 2005 until 2009. Malignant tumors group consisted of 71 patients with ages between 20–80 years. Benign tumors group contained 83 patients with ages between 17–55 years. Other tumors group comprised 100 patients with ages between 16–70 years.The data was summarized by using tables and graphs. Figure (1) demonstrated, a 95% confidence interval for the survival time for each group by remission status.



Figu

(*1 = Patient is Still in Remission,* 0 = *Censored*)

2. *Kaplan–Meier Survival Analysis*

The Kaplan–Meier estimator (*K–M*) is a non–parametric estimator which december be used to estimate the survival distribution function from censored data.The Kaplan–Meier estimator is also called *product–limit*

estimator (*PL*) because of its typical product structure[1].

The estimator is similar to the actuarial estimator except that the lengths of the intervals I_j are variable. In fact, let t_j, the right endpoint of I_j, be the jth ordered censored or uncensored observation.We observe

the pairs $(Y_1, \delta_1), \dots, (Y_n, \delta_n)$. For now, assume no ties. Let $0 \le Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)} < \infty$ be the order statistics of $Y_1, Y_2, \dots, Y_n$, and with an abuse of notation, define $\delta_{(j)}$ to be the value of δ associated with $Y_{(j)}$, that is, $\delta_{(j)} = \delta_j$ when $Y_{(j)} = Y_j$. Note that $\delta_{(1)}, \dots, \delta_{(n)}$ are not ordered. Let $\mathcal{R}(t)$ denote the *risk set* at time t, which is the set of subjects still alive at time $t-$, and let

$$n_j = number\ of\ subjects\ in \mathcal{R}(Y_{(j)}) = number\ of\ alive at time Y_{(j)}-$$

$$d_j = number\ of\ subjects\ died at time Y_{(j)}$$

$$p_j = P_r(surviving through I_j \setminus alive at beginning of I_j) = P_r(T > t_j \setminus T > t_{j-1})$$

From the estimates

$$\hat{q}_j = \frac{d_j}{n_j}$$

$$\hat{p}_j = 1 - \hat{q}_j = \begin{cases} 1 - \frac{d_j}{n_j} if \delta_{(j)} = 1 \ (uncensored) \\ 1 \qquad if \delta_{(j)} = 0 \ \ (censored) \end{cases}$$

$$\hat{S}(t) = \prod_{y_{(j)} \le t} \hat{p}_j = \prod_{j: y_{(j)} \le t} \left(1 - \frac{d_j}{n_j}\right) \qquad \cdots (1)$$

If all Y_j are different, then each $d_j = 1$ and $n_j = n - rank(Y_j) + 1$, and in this case, the product–limit estimate when no ties are present is

$$\hat{S}(t) = \prod_{y_{(j)} \le t} \left(1 - \frac{1}{n_j}\right) = \prod_{y_{(j)} \le t} \left(1 - \frac{1}{n_j}\right)^{\delta_{(j)}} = \prod_{y_{(j)} \le t} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} \qquad \cdots (2)$$

The variance of the estimator is given by:

$$\widehat{Var}\left(\hat{S}(t)\right) \approx \left(\hat{S}(t)\right)^2 \sum_{j: t_{(j)} < t} \frac{\delta'_{(j)} d_j}{n_j (n_j - d_j)} \qquad \cdots (3)$$

This is known as *Greenwood's formula*[6].

The Log–rank Test

The log–rank test is a non–parametric method for testing the null hypothesis that the groups being compared are samples from the same population as regards survival experience.The first step is to arrange the survival times, both observed and censored. Suppose, for illustration, that there are two groups, A and B. For each minute with a failure we calculate the numbers at risk in each group (r_A and r_B) and the numbers of observed failures (f_A and f_B). If at time t_j in groups A and B,

respectively, then the data can be arranged in a 2×2 table as follows[3]:

	Group A	*Group B*	*Total*
Died	f_A	f_B	f
Survived	$r_A - f_A$	$r_B - f_B$	$r - f$
Total	r_A	r_B	r

Except for tied survival times, $f = 1$ and each of f_A and f_B is 0 or 1. Note also that if a subject is censored at t_j then that subject is considered at risk at that time and so included in r.

On the null hypothesis that the risk of death is the same in the two groups, then we would expect the number of deaths at any time to be distributed between the two groups in proportion to the numbers at risk. That is,

$$E(f_A) = \frac{r_A f}{r} \quad \cdots (4)$$

$$Var(f_A) = \frac{r_A r_B f(r-f)}{r^2(r-1)} \cdots (5)$$

Summing over all times of death, t_j, gives

$$\left.\begin{aligned} O_A &= \sum f_A \\ E_A &= \sum E(f_A) \\ V_A &= \sum Var(f_A) \end{aligned}\right\} \quad \cdots (6)$$

Similar sums can be obtained for group B and it follows from (4) that$E_A + E_B = O_A + O_B$.A test statistic for the equivalence of the death rates in the two groups is

$$\chi_1^2 = \frac{(O_A - E_A)^2}{V_A} \cdots (7)$$

Which is approximately a χ_1^2. The log–rank statistic approaches to chi–square distribution with one degree of freedom[1]. The hazard ratio sampling variability are given by

$$h = exp\left(\frac{O_A - E_A}{V_A}\right) \quad \cdots (8)$$

$$SE[\ln(h)] = \sqrt{\frac{1}{V_A}} \quad \cdots (9)$$

3. Results

In this study, we chooses simple random sample of the patients were analyzed with using both descriptive and inferential statistics. The results of Kaplan–Meier method is analyze by (*SPSS*) statistical packages were used to analyze the data. And the results are introduced and tabulated in following Tables for applying Kaplan–Meier method to breast cancer data.

Table (1) displays tumor diagnosis, total number of patients for each diagnostic groups,

patients experienced event, and censored patients. It was noted that number of events for malignant and benign tumors group were similar, while other tumors group had the highest number of events

Table*(1):Summary of Remission Status*

Diagnosis	*Total No.*	*No. of event*	*Censored*	
			No.	*Percent (%)*
Malignant Tumor	71	43	28	39.4
Benign Tumor	83	61	22	26.5
Other Tumor	100	79	21	21.0
Overall	254	183	71	28.0

The means and medians for survival time table offers a quick numerical comparison of the "typical" times to effect for each of the tumors. Since there is a lot of overlap in the confidence intervals, it is likely that there is much difference in the "average" survival time, shows in Tables (2) and (3).

***Table*(2):** *Median for Survival Time*

Diagnosis	*Median*			
	Estimate	*Standard Error*	*95% Confidence Interval*	
			Lower Bound	*Upper Bound*
Malignant Tumor	7.000	1.201	4.646	9.354
Benign Tumor	84.000	17.330	50.034	117.966
Other Tumor	34.000	24.000	0.000	81.040
Overall	45.000	13.962	17.635	72.365

The mean of the survival times for each groups was computed. For malignant tumor group it was 267.365, for benign tumor group 368.238, and 266.770 for other tumors group. Since 39.4% of the times in malignant tumor group are censored, the true mean survival time for that group, in reality, might be higher (perhaps, considerably so) than 267.365. The true mean survival time for benign tumor group and other tumors group were also likely higher than the computed 368.238, and 266.770 respectively, but with 26.5%

censored time for benign tumor and 21% for other tumors. We did not expect as great a difference between the calculated mean and the true mean for both groups. Thus, we see that we had still another indication that the survival experience of benign tumor group is more favorable than the survival experience of malignant tumor group and other tumors groups.

***Table* (3):** *Mean for Survival Time*

Diagnosis	*Mean*[a]			
	Estimate	*Standard Error*	*95% Confidence Interval*	
			Lower Bound	*Upper Bound*
Malignant Tumor	267.365	72.180	125.893	408.838
Benign Tumor	368.238	57.928	254.698	481.777
Other Tumor	266.770	40.185	188.007	345.533
Overall	344.203	36.128	273.392	415.013

a. Estimation is limited to the largest survival time if it is censored.

These observation strongly suggest that the survival experience of patients with benign tumors was far more favorable than that of patients with malignant tumors and other tumors.The results of applying the log–rank test in this case were:

Table* (4):Group Comparisons by Log–rank Test*

Log–rank (Mantel–Cox)	*Chi-Square*	*DF*	*significant*
Malignant,Benign,& Other	5.487	2	0.064
Malignant &Benign	5.572	1	0.018
Malignant &Other	2.398	1	0.121
Benign & Other	0.889	1	0.346

Test of Equality of Survival Distributions for the Different Levels of Diagnosis.

The graph also allowed us to represent visually the median survival time and survival rates representation of the life tables such as the 1–year survival rate.In Figure (2) shows that the horizontal axis shows the time to event. In this plot, drops the survival curve to reach to zero, and ascend the hazard curve to reach to above 1.5. While the vertical axis

shows the probability of survival and the cumulative hazard. Thus, any point on the survival curve shows the probability that a patient on a given diagnosis will not have experienced relief by that time. The plot for malignant tumor below that of benign tumor throughout most of the trial, which suggests that malignant tumor december give faster relief than benign tumor. To determine whether these differences are due to chance, look at the comparisons Tables above.

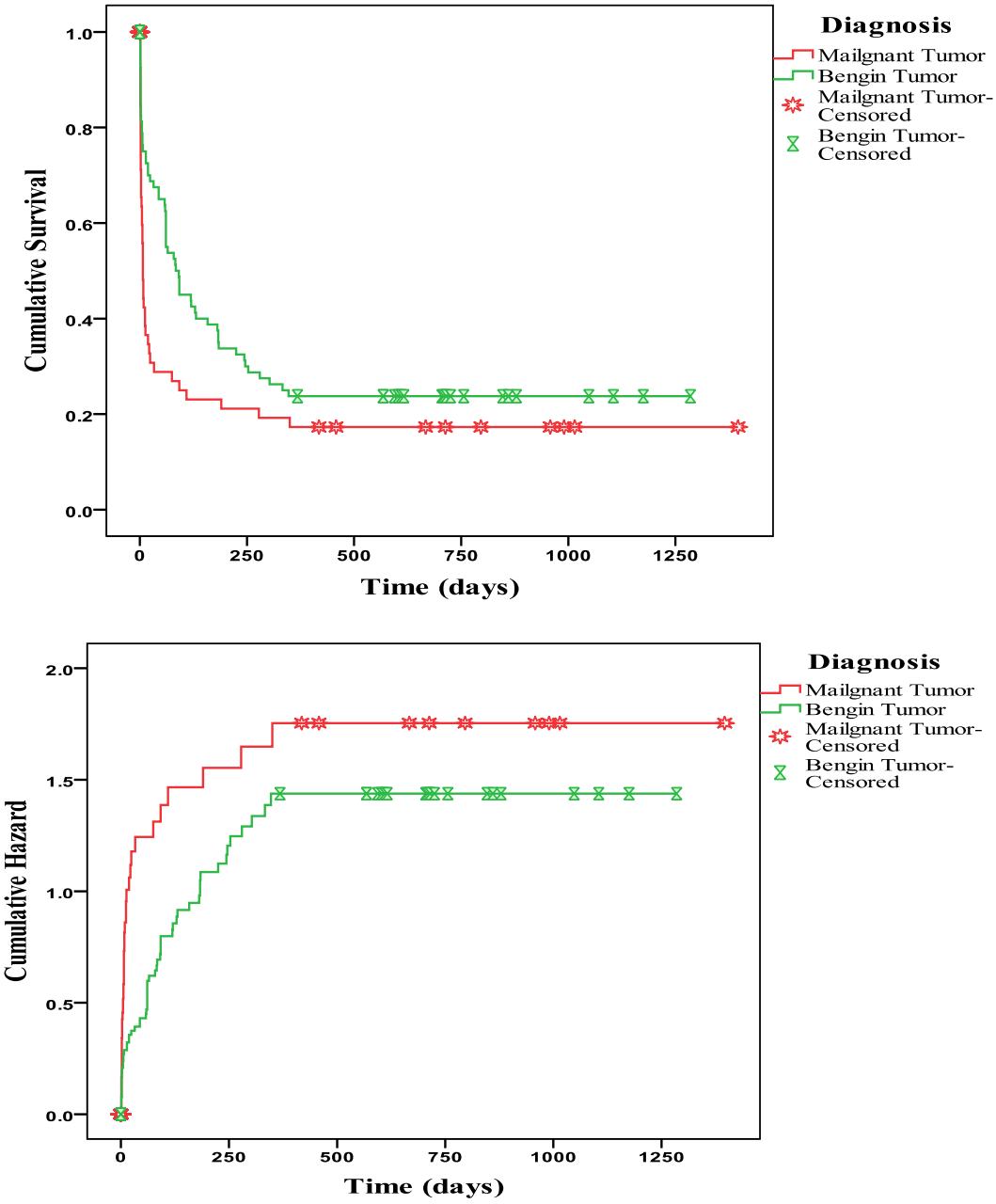


***Figure* (2):** *Comparison of Kaplan–Meier Survival Curve (Survival, Hazard) for Malignant and Benign Tumors Group.*

4. Conclusions

With the Kaplan–Meier survival analysis procedure, you have examined the distribution of time to effect for two or more different groups. The comparison tests show that there is a statistically significant differencein survival times $p < 5\%$between malignant and benign tumors group only.

5. References

[1] *Akbar, A., Pasha, G.R. and Naqvi, S.F.H.* (2009). "Properties of Kaplan-Meier Estimator: Group Comparison of Survival Curves". *European Journal of Scientific Research*,*Publish in, Inc.*,Vol. 32, No. 3, 391-397.

[2] *Al-A'bidy, M. A.* (2011) "Study and Analysis of Breast Cancer: Kaplan–Meier and Cox Proportional".*M.Sc. Thesis*,University of Kufa.

[3] *Altman, D.G.* (1991). "Practical Statistics for Medical Research". Chapman & Hall, New York.

[4] *Bland, JM. and Altman DG.* (1998). "Survival probabilities (the Kaplan-Meier method)". *BMJ*, Vol. 317, 1572, www.bmj.com.

[5] *Borgan, Ø.* (1997). "Three Contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, & Aalen-Johansen". P.O. Blindern, N-0316 Olso, Norway, University of Olso.

[6] *Jiezhi, Qi.* (2009). "Comparison of Proportional Hazards and Accelerated Failure Time Models". M.Sc. Thesis, University of Saskatchewan.

[7] *Rajaeefard, AR., Baneshi, MR, Talei, AR and Mehrabani, D.* (2009). "Survival Models in Breast Cancer Patients". *Iranian Red Crescent Medical Journal*, Vol. 11, No. 3, 295-300.

[8] *SAS/STAT® 9.2 User's Guide.* (2008). SAS Institute Inc., Cary, NC, USA.

[9] *Tovey, SM, Brown, S, Doughty, JC, Mallon, EA, Cooke, TG and Edwards* (2009). "Poor Survival Outcomes in HER2-Positive Breast Cancer Patients with low-grade, node-Negative Tumors". *British Journal of Cancer, Cancer Research UK*, Vol. 100, No. 5, 680-683.

[10] *Zino, S.M.W.* (2010). "Investigations into the expression of Sirtuins in breast cancer: *in vivo* and *in vitro* Studies". Ph.D. Thesis, University of Glasgow.

ملخص البحث:-

مُقدّر كابلان– مير من المُقدرات الشائعة جداً والذي يزوّد تقديرات أفضل لحساب الوسيط لحجوم عيّنات كبيرة إلى حدٍّ معقول.إنّ هدف هذا البحث هو بشكل رئيسي مهتمّ بدراسة وتحليل تقدير وقت البقاء للبيانات الحقيقية لمرضى سرطان الثدي في العراق