
Survival Analysis of Breast Tumor in Al-Najaf

Nazera Khalil Dakhil

Yahya Mahdi Al-mayali

Muna Abbas Mseer

College of Mathematics and Computer Sciences

University of Kufa

Abstract

Breast cancer is the most common type of tumor in Iraq. This research is mainly concerned with study analysis of breast tumor data. A representative random sample of patient suffering from tumor was collected from Al-Sadder hospital in Al-Najaf city. The data was then analyze using survival analysis technique.

Keywords: Breast tumor, Survival analysis, Kaplan-Meier, Iraq.

Introduction

Cancer is one of the five main causes of death in all societies. However Breast cancer in the Middle-East occurs in relatively young women and frequently presents as advanced disease. This study investigates breast tumor survival rates in the Al-Najaf city of Iraq. A random sample of data from Al-sadder hospital was selected. Al-Sadder hospital was build in 1984. It is the biggest and most importance teaching hospital in Al-Najaf city. It consisted of six floors with more than 13 specialized departments. It is a major referral hospital in Al-Najaf city and provided high standard health services to Al-Najaf province and its environs. The occurrence of breast cancer seem to be rising world-wide [1], and stable or declining

breast cancer in Western countries. These conflicting results are may be due to more effective therapies and to an early detection of small tumors. The increase in Western countries is well documented however in non-Western countries accurate statistics is difficult to collect.

In both the Middle-East and the West, carcinoma of the breast is the most common malignancy of women. In the Middle-East, breast cancer is often notice during the childbearing years [7]. Studies on breast cancer rates and risk factors in the Middle-East are limited, but Egypt is thought to have the highest rates in the Middle-East with the increase being primarily between 30 to 60 years of age [6].

A case control study from Iran investigators did, however, find a significantly increased risk among women who were never married and among women with a positive family history of breast cancer. They also found that the great majority of patients were diagnosed with advanced disease, and that many patients had a significant delay between their recognition of symptoms and their first medical consultation for the condition.

In this study we investigated the survival rates of breast tumor within Al-Najaf province, which is located south west of Baghdad. The aims of the study were to compare breast tumor survival rates, the malignant and benign tumor.

Data Analysis

The Directorate of Health in Al-Najaf serves as the ethics committee for Al-Sadder Hospital. The Directorate of Health gave permission for the research. The main source of the collected data was the specialized breast disease center in Al-Sadder hospital in Al-Najaf. Other sources of information were the statistics department in the same hospital.

A random sample of patients from breast cancer centre at Al-sadder hospital was reviewed. The data consisted of 59 patient records, that contained address, education, occupation, age, weight, date of the first consultation, and last date, of follow up. All the cases including in the present study were diagnosed as either malignant or benign lesions. Patients were followed up from a minimum of one year and maximum 5 years. In survival analysis, follow up periods were calculated from the first consultation with surgeon. Data were entered and coded. It was checked thoroughly then categorized. Occupation unemployed, and employed. Education was categorized into higher, further, secondary, primary, and illiteracy. The data was summarized using tables and graphs. A number and percentages was used to summaries categorical data and descriptive statistics for continuous data.

A survival analysis was used to compare the rates of patients with malignant tumor with benign one. SPSS statistical packages were used to analyze the data. Two tails hypothesis was used with 5% and 1% level of significant. Survival curves were generated according to Kaplan-Meier method and probabilities difference in survival were tested by long-rank test.

Survival analysis

This is a set of statistical techniques used when the primary outcome of interest is the time between patients' entry into a study and a subsequent outcome. In medical applications this outcome is usually death, but not always. *For example*, the outcome may be time to recurrence. [8]

Censoring

Suppose a study involves n patients, with survival times $T_1, \dots, T_n$. Each patient is associated with a censoring time C_i and the observations are

$$Y_i = \begin{cases} T_i & T_i < C_i \\ C_i & T_i \geq C_i \end{cases}$$

It is always known whether an observation is censored $Y_i = C_i$ or uncensored $Y_i < C_i$. [5]

A survival time is censored if all that is known is that it began or ended within some particular interval of time, and thus the total spell length (from entry time until transition) is not known exactly. We may distinguish the following types of censoring:

- *Right censoring*: at the time of observation, the relevant event (transition out of the current state) had not yet occurred (the spell end date is un-known), and so the total length of time between entry to and exit from the state is unknown. Given entry at time 0 and observation at time t , we only know that the completed spell is of length $T > t$.

- *Left censoring*: the case when the start date of the spell was not observed, so again the exact length of the spell (whether completed or incomplete) is not known. Note that this is the definition of left censoring most commonly used by social scientists. (Be aware that biostatisticians typically use a different definition: to them, left-censored data are those for which it is known that exit from the state occurred at some time before the observation date, but it is not known exactly when. See e.g. Klein and Moeschberger, 1997.) [3]

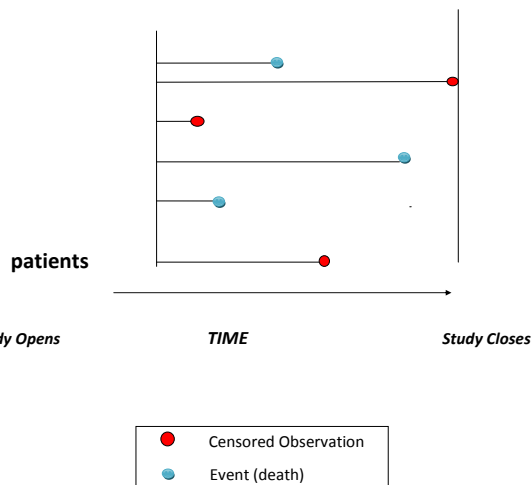


Illustration of survival data

Probability Functions [4]

If the probability density function is $f(t)$ and the cumulative probability function is $F(t)$, then the survivor function is

$$S(t) = 1 - F(t) \quad (1)$$

And the hazard function

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\log S(t)}{dt} \quad (2)$$

This is the rate or intensity of the point processes of the previous chapter. Then, we have

$$S(t) = \exp \left[-\int_0^t h(u) du \right] \quad (3)$$

$$f(t) = h(t) \exp \left[-\int_0^t h(u) du \right] \quad (4)$$

Where $\int_0^t h(u) du$ is called the *integrated hazard* or *intensity*.

Suppose that I_i is a code or indicator variable for censoring, with $I_i = 1$ if the observation i is completely observed and $I_i = 0$ if it is censored. Then, the probability for a sample of n individuals will be approximately (because the density assumes that one can actually observe in continuous time) proportional to

$$\prod [f(t_i)]^{I_i} [S(t_i)]^{1-I_i} \quad (5)$$

And a likelihood function can be derived from this. In most cases, this does not yield a generalized linear model.

Kaplan-Meier Estimates[2]

Let $t_{(1)} < t_{(2)} < \dots < t_{(m)}$ denote the distinct times in which an event was observed, d_i the number of events that occurred at time $t_{(i)}$, and r_i the size of the risk set at time $t_{(i)}$. The Kaplan-Meier estimate for a survival function, also called *product-limit estimate*, is given by:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_{(1)} \\ \prod_{t_{(i)} \leq t} \left[1 - \frac{d_i}{r_i}\right] & \text{if } t \geq t_{(1)} \end{cases} \quad (6)$$

The Kaplan-Meier estimate $\hat{S}(t)$ is a right-continuous step function with jumps in the event times. Censoring times effect the estimate only by reducing the risk set for next event, thereby increasing the height of the next jump.

In the presence of censoring, Greenwood (1926) suggested the following estimate for the variance of the Kaplan-Meier estimate:

$$\widehat{Var}(t) = \hat{S}(t)^2 \sum_{t_{(i)} \leq t} \frac{d_i}{r_i(r_i - d_i)} \quad (7)$$

The Kaplan-Meier estimate $\hat{S}(t)$ is asymptotically normally distributed. This leads to the following point-wise confidence intervals for the survival function, $\hat{S}(t)$,

$$\hat{S}(t) \pm z_{1-\alpha/2} \widehat{Var}(t)^{1/2} \quad (8)$$

Where $(1 - \alpha)$ is the coverage probability, z_p denotes the $p \times 100th$ percentile of the standard normal distribution, and $\widehat{Var}(t)$ is Greenwood's estimate of the variance of $\hat{S}(t)$, given in formula (7). Note that Greenwood's estimate tends to slightly underestimate the true variance, so that the true coverage probability of the confidence intervals might be somewhat smaller than stated.

The Log-rank test [2]

To compare two survival curves produced from two groups A and B, we use the rather curiously named log rank test, so called because it can be shown to be related to a test that uses the logarithms of the ranks of the data.

As for the Kaplan-Meier survival curve, we now consider each event in turn, starting at time $t = 0$. At each event (death) at time t_i we consider the total number alive (r_i) and the total number still alive in group A (r_{Ai}) up to that point. If we had a total of d_i events at time t_i , then, under the null hypothesis, we consider what proportion of these would have been expected in group A. Clearly, the more people at risk in one group the more deaths (under the null hypothesis) we would expect. Thus the expected number of events in A is

$$E_{Ai} = \frac{r_{Ai} d_i}{r_i} \quad (9)$$

The effect of the censored observations is to reduce the numbers at risk, but they do not contribute to the expected numbers.

Finally, we add the total number of expected events in group A ,

$$E_A = \sum E_{Ai} \quad (10)$$

If the total number of events in group B is E_B , we can deduce E_B from $E_B = n - E_A$. We do not calculate the expected number beyond the last event, in this case at time 42 months. Also, we would stop calculating the expected values if any survival times greater than the point we were at were found in one group only.

Now, to test the null hypothesis of equal risk in the two groups we compute

$$X^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \quad (11)$$

where O_A and O_B are the total number of events in groups A and B . We compare X^2 to a χ^2 distribution with one degree of freedom (one, because we have two groups and one constraint, namely that the total expected events must equal the total observed).

The relative risk can be estimated by

$$R = \frac{(O_A/E_A)}{(O_B/E_B)} \quad (12)$$

The standard error of the log risk is given by

$$SE(\log(R)) = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}} \quad (13)$$

Results and Conclusions

The random sample of patients were analyzed using both descriptive and inferential statistics.

The data consisted of 69 women, of which 40 were diagnosed with malignant tumor and 19 with benign tumor. 98% of patients with benign tumor are living in Al-Najaf and 2% were missing. 90% of patients with malignant tumor are living in Al-Najaf and 10% are living in nearby province. 100% of patients with malignant were unemployed compared with 98% of patients with benign tumor, and only 2% of patients with benign tumor were employed.

Level of education for patients with benign tumor were distributed as follow: 38% illiteracy, 43% primary, 10% secondary, and 9% higher, while level of education for patients with malignant tumor were distributed as follow: 53% illiteracy, 32% primary, 5% secondary, and 10% higher.

Table 1: Descriptive statistics for patients with begin tumor

	N	Minimum	Maximum	Mean	Std. Deviation
Time in month	19	.00	54.63	11.7532	19.67778
Age	19	26	70	52.68	11.494
Weight	8	56	100	70.66	15.235

Table 1 showed that 18 patients had a valid age and one missing. Only 8 had a valid weight. The mean of age was 52 years and about 71 kg weight. The mean for survival time in days was 354.

Table 2: Descriptive statistics for patients with malignant tumor

	N	Minimum	Maximum	Mean	Std. Deviation
Time in month	40	.03	46.40	9.1997	12.96510
Age	40	9	55	36.97	9.512
Weight	20	60	98	76.08	9.941

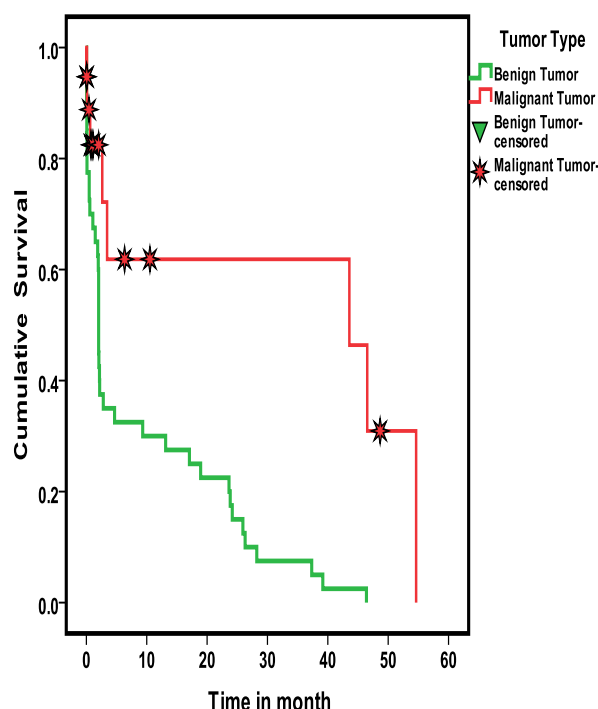
Patients with malignant tumor were 40. The mean of age was 37 years and about 76 kg weight. The mean for survival time in days was 171.

The mean age for patients suffered from benign were higher than patients with malignant tumor. While the mean weight for patients suffered from malignant were higher than patients with benign tumor. This was indicated that patients with malignant tumor were younger the patients with benign.

Table 3: Means and Medians for survival time

Dignosis	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Benign Tumor	9.200	2.050	5.182	13.21	2.030	.031	1.968	2.092
Malignant Tumor	31.499	7.521	16.757	46.24	43.570	23.980	.000	90.571
Overall	13.220	2.462	8.395	18.04	2.170	.257	1.666	2.674

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



Kaplan –Meir survival analysis of 59 patients who were followed up showed a median survival of 1307 days for malignant and 61 days for benign, the difference was found to be highly significant ($\chi^2 = 10.44$, $df = 1$, $P = 0.001$) at 1% level of significance (figure 2).

References

- [1] Althius MD, Dozier JM, Anderson WF, Devessa SS, Brinton LA: Global trends in breast cancer incidence and mortality 1973-1997. *Int J Epidemiol* 2005, 34:405-412.
- [2] Hardle W., Zdenek H, and Klinke S. *XploRe Applications Guide* (e book) 2003.

-
- [3] Jenkins S.P. Survival Analysis, University of Essex, Uk.2005.
- [4] Lindsey J.K. Applying Generalized Linear Models 1997.
- [5] Loader C. Local Regression and Likelihood. Springer 1999.
- [6] McCredie M, Coates M, Grulich A: Cancer incidence in migrants to New South Wales (Australia) from the Middle-East. 1972-91. *Cancer causes and control* 1994, 5:414-421.
- [7] Ries LAG, Melbert D, Krapcho M, Stinchcomb DG, Howlader N, Horner MJ, Mariotto A, Miller BA, Feuer EJ, Altekruse SF, Lewis DR, Clegg L, Eisner MP, Reichman M, Edwards BK, (eds): [http://seer.cancer.gov/csr/1975_2005/] webcite *SEER Cancer Statistics Review, 1975-2005*. National Cancer Institute. Bethesda, MD; 2008.
- [8] Smith F.G., and Smith J.E. Clinical Research. Bios Scientific Publishers Limited 2003.
- [9] Swinscow T.D., and Campbel M.J. Statistical at Square One. BMJ Books 2002.

تحليل البقاء لأورام الثدي في محافظة النجف

منى عباس مسير

أم. د. يحيى مهدي هادي الميالي

م. د. نزيهة خليل داخل

كلية الرياضيات وعلوم الحاسوب- جامعة الكوفة

ملخص البحث

أورام الثدي هي ربما أكثر الأورام انتشاراً في العراق. يتناول هذا البحث دراسة وتحليل بيانات أورام الثدي. جمعت العينة العشوائية من المرضى المصابين بأورام الثدي والذين يتلقون علاجهم في مستشفى الصدر التعليمي في محافظة النجف الأشرف. وقد تم تحليل البيانات بطريقة تحليل البقاء.