Pakistan Vet. J., 2007, 27(4): 194-198.

SURVIVAL ANALYSIS OF CANCER PATIENTS USING PARAMETRIC AND NON-PARAMETRIC APPROACHES

M. AKRAM, M. AMAN ULLAH AND R. TAJ

Department of Statistics, Bahauddin Zakariya University Multan, Pakistan

ABSTRACT

Exploring the health related quality of life is usually the focus of the survival studies. Using the health data of cancer registry in Multan, Pakistan, an investigation about the survival pattern of cancer patients was explored, using the non-parametric and parametric modeling strategies. The Kaplan-Meier method and Weibull model based on Anderson-Darling test were applied to the real life time data. Findings suggested different sex-superiority of survival pattern among different groups of cancer patients. Interestingly, Kaplan-Meier and Weibul model provided a very close estimate of the survival function and other characteristics of interest.

Key words: Anderson-Darling test, cancer patients, hazard function, Kaplan-Meier estimate, survival function.

INTRODUCTION

At an individual level, diagnosis of cancer is regarded as a human tragedy. At the level of society, cancer is one of the major chronic diseases, causing a notable amount of health administrative costs. Prognosis and possible cure from cancer are thus important measures of lifetimes which can be assessed by analyzing the survival of cancer patients.

Different statistical approaches are used for analyzing the cancer survival data. The results of survival analysis for cancer patients have been widely presented and reported for different human sub populations of the globe (Woolson, 1981; Kardaun, 1983; Beadle *et al*., 1984; Sedmak *et al*., 1989). However, very few survival results at national level are available for the population of Pakistan (Khan *et al*., 2004). The statistical evidence about the survival of the cancer patients in the region of the Southern Punjab (Pakistan) is not available in the literature. McGarty (1974) has mentioned that for adopting any suitable statistical technique for analyzing survival data, it should be assumed that the statistical model embodies the evaluation of some natural process with the believe that the model is a useful approximation of the real process. Several approaches have been proposed in the literature by Leung *et al.* (1997) and Little and Rubin (2002) for analyzing the survival data.

The main objectives of this study were (i) to estimate the survival function S(t), using the standard Kaplan-Meier estimator, (ii) to estimate the cumulative hazard function H(t), using the Nelson-Aalen estimator and (iii) to fit an appropriate parametric lifetime model based on Anderson-Darling goodness of fit test.

MATERIALS AND METHODS

The relevant lifetime data on the patients of cancer in accord with the Nishtar Hospital Multan (Pakistan) was selected. This hospital receives its patients from a wide area in the limits of Southern Punjab. In this study, a retrospective simple random sample design was used; the lifetime data on 202 male and 145 female patients of cancer belonging to different classes was selected. These 347 patients of cancer were treated in Nishtar Hospital Multan during January, 1997 to December, 2001. The registration time was January 1, 1997 to June 30, 1997.

Assumption and notations

In this study the generalized type-I censoring was considered. For more convenience, the censoring was due to the following reasons: (i) A patient emigrated out of the study area was impossible to follow. (ii) An individual survived past the end of the study period. (iii) The censoring was non-informative.

For the representation of the data considered in this study, each individual had its own specific lifetime which was rescaled at starting time to $t_0 = 0$ (Klein and Moeschberger, 1997). T was taken as a non-negative random variable, the time until the event of interest (death) due to cancer occurred. It was assumed to be independently and identically distributed with probability density function f(t), the survival function S(t) and hazard function h(t). C_r was the fixed right censoring time; T and C_r were assumed to be independent. The exact lifetime of an individual was known if and only if T was less than or equal to C_r. Pairs of random variables conveniently represented the data, (X, δ) where δ was the censoring indicator and X was equal to T , if the lifetime was observed , and to C_r if it was censored and X=min(T, C_r).

Parametric approach

Considering the lifetime parametric model as the useful approximation of the real process, three lifetime models viz Exponential, Weibull and Gamma distribution were considered. The Anderson-Darling test, which makes the use of these specific lifetime distributions in calculating critical values, was defined with

H_0: The data followed a specified parametric lifetime model.

Ha: The data did not follow the specified lifetime model.

The Anderson and Darling (1954) test statistic was

$$A^2 = -\sum_{i=1}^{n} \frac{(2i-1)}{n}\left[\log F(Y_i) + \log(1-F(Y_{n+1-i}))\right] - n,$$

where F was the cumulative distribution function of the specified distribution, Y_i was the ordered data and n was the number of observations. The test was a one-sided test and the hypothesis that the distribution of a specific form was rejected if the test statistic 'A' was greater than the critical value. From the class of specified lifetime distributions, the parametric lifetime model, which one has the minimum Anderson-Darling (adjusted) value, gave the better fit.

Nonparametric approach

Cox and Oakes (1984) and Kalbfleisch and Prentice (2002) presented a nonparametric approach to estimate survival function using standard Kaplan Meier (KM) technique (Kaplan and Meier, 1958). There were D distinct times with $t_1 < t_2 < \ldots t_D$, d_i deaths or events occurred at time t_i and Y_i were the number of individuals who were at risk at time t_i. The KM estimator was defined as for all values of t in the range where there was data:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i < t}(1 - \frac{d_i}{Y_i}) & \text{if } t_i < t \end{cases}$$

It was obvious from KM estimator, for $t < t_1$, $\hat{S}(t) = 1$ and when $Y_i = d_i$, then $\hat{S}(t) = 0$ for $t \geq t_i$.

Cox and Oakes (1984) also established the variance of the KM estimator using Greenwood's relation as: $\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i\left(Y_i - d_i\right)}$

Moreover, KM estimator was also used to estimate the cumulative hazard function; $\hat{H}(t) = -\ln[\hat{S}(t)]$

The Nelson Aalen (NA) estimator of the cumulative hazard rate was defined up to the largest observed time on the study (see Aalen, 1978):

$$\hat{H}(t) = \begin{cases} 0 & if\ t \leq t_1 \\ \sum_{t_1 \leq t} \frac{d_i}{Y_i} & if\ t_1 \leq t \end{cases}$$

The estimated variance of the NA estimator was given by

$$\hat{\sigma}_H^2(t) = \sum_{t_1 \leq t} \frac{d_i}{Y_i^2}$$

RESULTS AND DISCUSSION

Cancer is an important public health concern throughout the world (Greenlee *et al.,* 2000). Torey and Broom (2007) reported that cancer is the second leading cause of death in Pakistan. The study was inspected by observing the survival times given for each gender group of different classes of cancer in the data set. Love (1991) noted in his study that in low-income and minority populations, fatalism about cancer and negativism about cancer therapy were widespread. In the present study, the two descriptive measures, average survival time $(\overline{T})$ and the average hazard rate $(\overline{h})$ were used for the overall comparison by ignoring the phenomenon of the censorship, which are given in Table 1. In order to calculate these measures, survival time of each patient for each group was used.

On the whole, it appeared from Table 1 that the female group survived more than the male group based on the two descriptive measures. These descriptive measures did not compare the two groups at different time points in time of follow up; however, such a visual comparison of gender survival by using the standard non-parametric KM and NA methods for the group of bone tumor patients is presented in Fig. 1 and Fig. 2, respectively.

It is apparent from Fig. 1 that female group consistently lay above that for male group particularly upto 75 years of age. This difference indicates that female patients are the better survivors. Fig. 2 is the plot of the cumulative hazard rate for bone tumor

Table 1: Descriptive measures of survival time and hazard rate

Cancer group	Average survival time in years ($\bar{T}$)		Average hazard rate ($\bar{h}$)	
	Males	Females	Males	Females
Bone tumor	40.5	50.4	0.0120	0.0066
Brain tumor	50.2	51.9	0.0080	0.0092
Lung cancer	45.5	44.7	0.0047	0.0050
Leukemia	46.6	52.0	0.0068	0.0068
Liver cancer	42.4	51.9	0.0066	0.0090
Oran cancer	49.8	50.5	0.0100	0.0099
Overall	45.8	50.2	0.0080	0.0077

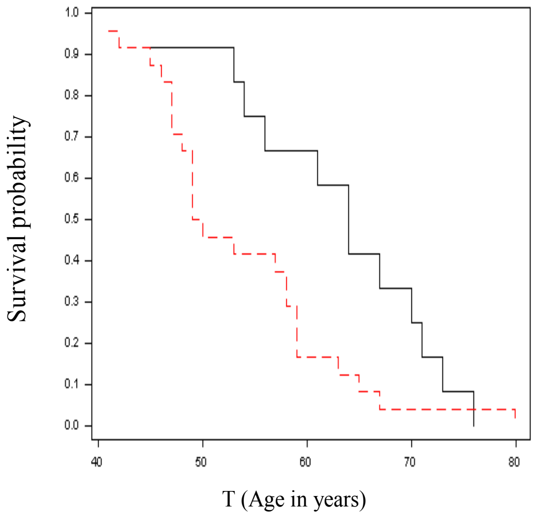

Fig. 1: Comparison of survival function of male and female bone tumor patients by using KM estimator.

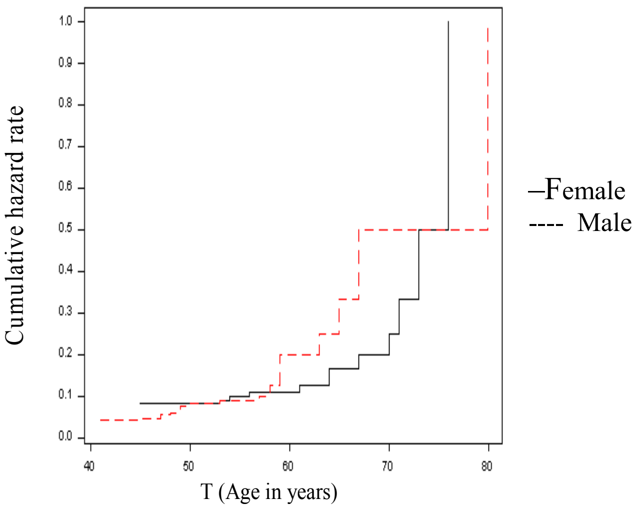

Fig. 2: Comparison of cumulative hazard function of male bone tumor and female patients by using NA estimator.

patients, which also shows that female prognosis of survival were better than their male counterparts. In a similar fashion, the survival prognosis about the remaining types of the cancer patients can also be determined.

To explore the statistical significance of gender survival, Log-rank and Wilcoxon test were used. In Table 2, the p-values for both tests were near to zero which provided the strong statistical evidence that males were dying faster than females. The empirical results of descriptive characteristics due to non-parametric KM and best fitted Parametric Weibull model approaches are presented in Tables 3 and 4 respectively. Table 3 shows that female patients had greater mean survival time (MST) of 64.11 years than 53.63 years for males. Alidina *et al.* (2004) also used the Kaplan-Meier approach to estimate the mean survival time for the esophageal cancer patients in Pakistan. They estimated the mean age of 56 years in 59 percent male and 41 percent female patients, while in this study, 58 percent males and 42 percent females had mean age of 42.4 and 51 years, respectively. Table 3 also shows that the estimate of median survival time for males was 49 years progressing to 64 years for female patients, which again confirms the survival superiority of females.

Table 2: Statistical significance tests of gender survival

Test	Chi-Square	Degree of freedom	P-Value
Log-Rank	8.7362	1	0.0031
Wilcoxon	12.3546	1	0.0004

From the parametric point of view, the Weibull distribution seemed to be the best fitted life time model

Table 3: Descriptive characteristics of bone tumor patients by Kaplan-Meier procedure

MST (years)		Standard error		95% Normal CI				Median		Q_1		Q_3	
				Lower		Upper							
F	M	F	M	F	M	F	M	F	M	F	M	F	M
64.11	53.63	2.29	1.86	59.62	49.97	68.58	57.28	64.00	49.00	56.00	47.00	71.00	59.00

M= Male, F= Female, MST= Mean Survival Time, Q_1= First Quartile, Q_3= Third Quartile

Table 4: Parameter estimates and major characteristics of interest of Weibull Distribution for bone tumor patients

Parameter	Estimate		Standard error		95.0% Normal CI			
	Male	Female	Male	Female	Lower male	Lower female	Upper male	Upper female
Shape	8.53	5.73	1.99	0.82	5.40	4.34	13.48	7.57
Scale	66.63	57.54	2.37	2.18	62.14	53.42	71.45	61.97
Mean survival time	62.94	53.24	2.54	2.20	58.15	49.10	68.13	57.73
Standard deviation	8.79	10.76	1.75	1.24	5.95	8.59	13.00	13.49
Median	63.83	53.97	2.55	2.26	59.03	49.72	69.02	58.58
First Quartile	57.58	46.29	3.25	2.61	51.54	41.45	64.32	51.70
Third Quartile	69.23	60.91	2.35	2.19	64.78	56.76	73.99	65.36

Anderson-Darling (adjusted) female =1.04

Anderson-Darling (adjusted) male = 1.43

based on the lower values of Adjusted Anderson-Darling test for males and females for all classes of cancer. The parameter estimates alongwith the major characteristics of the distribution are shown in Table 4. Mean survival time (MST) of males (53.24 years) was lower than 62.94 years for females. The percentage deviations of MST for males and females were 0.73 and 1.84, which indicated a very close estimation of MST by using both approaches i.e. non-parametric and parametric. By observing the comparative graphs of survival function and hazard function based on KM estimator, NA estimator and Weibull lifetime model, it was observed that at young age survival rate of tumor patients was highest, while as age increased survival rate decreased.

The curve of probability density function in Fig. 3a describes the distribution of lifetime data. The probability plot (Fig. 3b) was used as a diagnostic tool to assess whether a particular distribution fitted on lifetime data. In Fig. 3b, the points fall very closer around the fitted line of Weibull distribution which indicates a better fit. Also Weibull distribution provided a better fit in connection with AD test statistic. The survival function given in Fig. 3c directly gives the median survival time of males as 54 years and 64 years for females, suggesting an equivalence of the results based on KM method. Fig. 3d shows the parametric cumulative hazard plot for the patients of bone tumor cancer.

In this study, the cancer related gender status was explored using the techniques of survival analysis. But the applicability of the results is limited in the context of Multan region only. Interestingly, these results are consistent with the natural gender survival that the females have the better predictive survival than males. The information generated in this article would help the policy makers about the relevant input for improving the quality of life of cancer patients.

REFERENCES

Aalen, O. O., 1978. Nonparametric inference for a family of counting processes. Ann. Stat., 6: 701-726.

Alidina, A., A. Gaffar, F. Hussain, M. Islam, I. Vaziri, I. Burney, A. Valimohd and W. Jafri, 2004. Survival data and prognostic factors seen in Pakistani patients with esophageal cancer. Ann. Oncol., 15: 118-122.

Anderson, T. W. and D. A. Darling, 1954. A test of goodness of fit. J. Amer. Stat. Assoc., 49: 765-769.

Beadle, G. F., J. R. Harris, B. silver, L. Botnick and S. A. H. Hellman, 1984. Cosmetic results following primary radiation therapy and adjuvant chemotherapy for early breast cancer. Cancer, 54: 2911-2918.

Cox, D. R. and D. Oakes, 1984. Analysis of Survival Data. Chapman and Hall, New York, USA.

Greenlee, R.T., M. B. Hill-Harmon, T. Murray and M. Thun, 2001. Cancer Statistics. Cancer J. Clin., 50: 7-33.

Kalbfleisch, J. D. and R. L. Prentice, 2002. The statistical analysis of failure time data. John Wiley and Sons Inc., New York, USA.

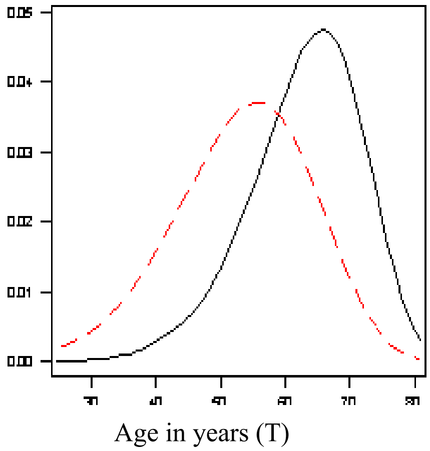

Fig. 3a: Probability density plot.

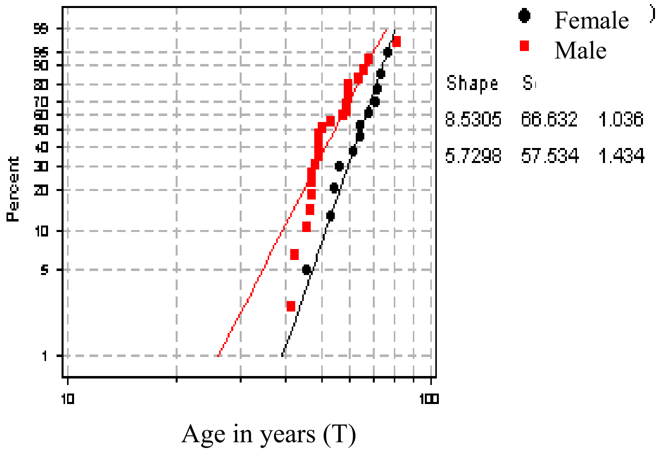

Fig. 3b: Weibull probability plot.

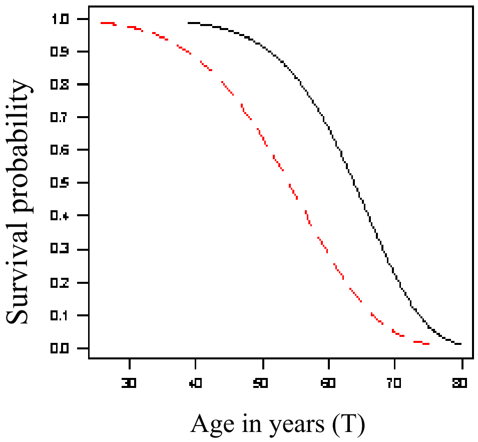

Fig. 3c: Parametric survival plot.

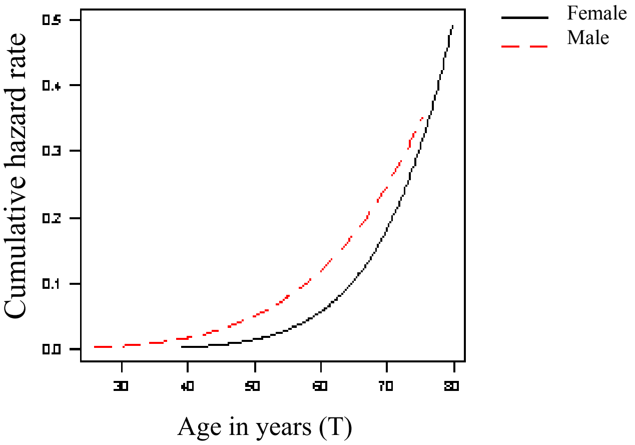

Fig. 3d: Parametric cumulative hazard plot.

Kaplan, E. L. and P. Meier, 1958. Nonparametric estimation from incomplete observations. J. Amer. Stat. Assoc., 53: 457-481.

Kardaun, O., 1983. Statistical analysis of male larynx-cancer patients: A case study. Statistical Nederlandica, 37: 103-126.

Khan, T. H., S. Iqbal, M. Abram and A. S. Warriach, 2004. The relationship of age, sex and marital status with the prevalence of cancer in the patients visiting Nishtar Hospital, Multan, Pakistan. Pakistan J. Zool., 36: 53-57.

Klein, J. P. and M. L. Moeschberger, 1997. Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag Inc., New York, USA.

Leung, K. M., R. M. Elashoff and A. A. Afifi, 1997. Censoring issues in survival analysis. Annual Review of Public Health, 18: 83-104.

Little, R. J. A. and D. B. Rubin, 2002. Statistical Analysis With Missing Data. John Wiley and Sons Inc., New York, USA.

Love, N., 1991. Why patients delay seeking care for cancer symptoms: what you can do about it. Cancer Eval., 89: 151-153.

McGarty, T. P., 1974. Stochastic System and State Estimation. John Wiley and sons Inc., New York, USA.

Sedmak, D. D., T. A. Meineke, D. S. Knechtges and J. Anderson, 1989. Prognostic significance of cytokeratin-positive breast cancer metastases. Modern Pathol., 2: 519-520.

Torey, P. and A. Broom, 2007. Cancer patient's negotiations of therapeutic options in Pakistan. Qualitative Hlth. Res., 17: 652-662.

Woolson, R. F., 1981. Rank test and a one-sample log rank test for comparing observed survival data to a standard population. Biometrics, 37: 687-696.