International
Journal
Of Advanced Research in Engineering& Management (IJAREM)
ISSN: 2456-2033 || PP. 08-12

Analysis of the Effects of Age Factor on Survival Time in Colorectal Cancer with Kruskal-Wallis Test

Odul Sanaroglu Buyruk, Kamil Alakus [2]
**(Department of Statistic, Ondokuz Mayıs University, Turkey)*
*** (Department of Statistic, Ondokuz Mayıs University, Turkey)*

Abstract: The most important step after a person's disease is diagnosed is starting the appropriate treatment. If this disease is terminal, the phenomenon of failure in treatment is death. Throughout the research, the time that passes from the diagnosis date of patients to the last observation or to the phenomenon of failure, if it has occurred, is the survival time. In such cases, survival analysis is used to estimate the patient's survival time or the probability of death. Survival analysis is a flexible method that can also include censored data. The present study discusses colorectal cancer, which is one of the most common cancer types. Since colorectal cancer is most commonly seen in middle aged and older individuals, the association between the disease and age was examined. This study was conducted with Kruskal-Wallis test, which is a survival analysis method, and the results were presented.

Keywords: colorectal cancer, censored data, kruskal-wallis test, survival analysis

I. INTRODUCTION

Survival analysis is an integration of methods which are used to estimate the longevity, probability of death, mean lifespan and average length of life by considering various factors within the treatments an individual receives after s/he is diagnosed with a disease. Survival analysis is also called life analysis or reliability analysis in some sources.

Survival analysis is used not only in medicine, but also in areas such as actuary and engineering. The phenomenon of "death" is considered as the same with "impairment" or "breakdown" in engineering and it is applied based on the assumption of survival analysis. The fact that the examined situation is seen in the subject means "failure". In other words, it means disease or death in living beings and breakdown for non-living things [1].

Survival analysis is a very commonly used type of analysis in medicine, especially in the analysis of cancer data. Survival analysis is used when one wants to estimate the probability of survival in cases such as general survival, disease-free survival, 5-year-long survival, etc.

The difference between non-parametric survival methods and other non-parametric tests is the fact that they can also work with censored data. In a study conducted in the field of medicine, a patient observed in a research can leave before the research ends, can die due to a different reason or can be still alive when the research ends. In this case, the data obtained is "censored". If the expected case occurs (death, etc.), the data obtained is "uncensored".

Cancer can be defined as the uncontrolled division and proliferation of cells in any part of the body. These developing cells disrupt the functions of the organs in which they settle and threaten life. Cancer cells can spread to other organs in the body through blood and lymph and this is called metastasis.

Colorectal cancer is the cancer which starts at the colon or rectum part of the intestine. It can also be called colon or rectum cancer depending on the starting place; however, they are accepted to be in the same group since they show common features [2].

Colorectal cancer is the most commonly observed type of cancer in men and women in the world [3]. More people will be diagnosed with colorectal cancer in the future; thus, colorectal cancer is a very important disease and more studies should be conducted on it.

II. METHODOLOGY

2.1. Functions Used In Survival Analysis

2.1.1. Survival (Life) Function

Probability distribution of survival time is called life or survival function. Survival function graph is called survival curve. Survival function expresses the general tendency of life data with a mathematical model. Survival function is a probability function and is shown with S(t) [4]. Given that t:any time and T: time of death; S(t)=P(T>t). S(t) function is also called cumulative life function. The estimated value of S(t) can be calculated as follows:

Sˆ(t)= (the number of individuals who survive longer than t)/(Total number of individuals)

2.1.2. Probability Density Function

f being a continuous function and T being a continuous variable that shows longevity, the limit value of the monitoring to fail within a small time interval shows the probability density function of T and the equation (1) is as given below [5].

$$f(t) = lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} . \qquad (1)$$

F(t) being the cumulative probability density function, $F(t) = \int_0^t f(x)\,dx;\ \ 0 < x < \infty$.

2.1.3. Hazard (Risk) Function

Hazard (Risk) Function expressed with h(t) shows the death risk of an individual living in t time, which may occur within a small time interval (t+Δt). Hazard Function is accepted as the criterion of failure.
Hazard Function is calculated as given in Equation (2).

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t / T \geq t)}{\Delta t} \qquad (2)$$

for continuous distributions, h(t) function meets the conditions of

$h(t) \geq 0,\qquad \int_t^{\infty} h(t)dt = \infty \qquad$ [6].

2.2. Non-Parametric Survival Function Estimates

d_i being the number of deaths at t_i time and, n_i being the number of individuals under risk at t_i time, various estimation methods are as given in Table. 1.

Table.1: Survival Function Estimators [7].

Researcher (s)	Formula
Kaplan - Meier (1958)	$S_j^{KM} = \prod_{i=1}^{j} \left(\frac{n_i - d_i}{n_i}\right)$
Altshuler (1970)	$S_j^{ALT} = \prod_{I=1}^{j} exp\left(-\frac{d_i}{n_i}\right)$
Prentice (1978)	$S_j^{*PREN} = \prod_{I=1}^{j} \left(\frac{n_i}{n_i + 1}\right)$
Prentice - Marek (1979)	$S_j^{*} = \prod_{i=1}^{j} \left(\frac{n_i - d_i + 1}{n_i + 1}\right)$
Harris - Albert (1991)	$S_j^{**} = \prod_{i=1}^{j} \left(\frac{n_i + d_i - 1}{n_i + d_i}\right)$

Fleming - Harrington (1991)	$S_j^{FH} = \exp\left(-\sum_{i=1}^{j} \frac{d_i}{n_i}\right)$
Moreau et al. (1992)	$S_j^{PREN} = \prod_{i=1}^{j}\left(\frac{n_i}{n_i + d_i}\right)$

2.3. Comparison of Survival Functions

In Life Analysis, Kruskal-Wallis test can be used when comparing the data in multiple groups.

2.3.1. Kruskal-Wallis Test

S_i , *total ordinal number of uncensored observations*

$$K = \sum_{i=1}^{r} X_i \ , \quad N = \sum_{i=1}^{r} n_i$$

$X_i \sim B(n_i, p),\ X_i$ *being independent ,*

$$E\left(S_i - X_i \frac{K+1}{2}\right) = 0$$

$$Var\left(S_i - X_i \frac{K+1}{2}\right) = \frac{n_i(N - n_i)p^2\{(N-2)p+3\}}{12}$$

$$Cov\left(S_i - X_i \frac{K+1}{2}, S_{i'} - X_{i'} \frac{K+1}{2}\right) = -\frac{n_i n_{i'}}{12}\{(N-2)p+3\}p^2, \qquad i \neq i'$$

and

$$Cov\left(S_i - X_i \frac{K+1}{2}, X_{i'} - n_{i'} \frac{K}{N}\right) = 0, \ \ \forall\, i, i'$$

If a random vector is thought of for :

$$Q' = \left[S_1 - X_1 \frac{K+1}{2}, \dots, S_{r-1} - X_{r-1} \frac{K+1}{2}, X_1 - n_1 p, \dots, X_1 - n_1 p\ \right]$$

Covariance matrix of Q ;

$$V = \begin{bmatrix} A & O \\ O & B \end{bmatrix}, \ \ B = p(1-p)D_2, \qquad D_2\ (rxr)\ diagonal\ matrix\ d_{ii} = n_i$$

$$A = \frac{p^2\{(N-2)p+3\}}{12}\,(N D_1 - \eta\eta)$$

$(r-1) \times (r-1)$ dimensioned D_1 being the left top bottom matrix of D_2 and $\eta^{T} = [n_1, \dots, n_{r-1}]$

$$G = Q^T V^{-1} Q$$

$$G = \frac{12}{np^2\{(N-2)p+3\}} \sum_{1}^{r} \frac{1}{n_i}\left(S_i - X_i \frac{K+1}{2}\right)^2 + \frac{1}{p(1-p)} \sum_{1}^{r} \frac{1}{n_i}(X_i - n_i p)^2$$

since p value is generally unknown, it is estimated with K/N [8] and test statistic is transformed into Equation (3).

$$G^* = \frac{12N^2}{K^2\{(N-2)K+3N\}} \Sigma_1^r \frac{1}{n_i}\left(S_i - X_i \frac{K+1}{2}\right)^2 + \frac{N^2}{K(N-K)} \Sigma_1^r \frac{1}{n_i}\left(X_i - n_i \frac{K}{N}\right)^2 \qquad (3)$$

The hypotheses to be tested here are made as;

H_0 : There are no differences between groups in terms of survival time.
H_1 : Survival time of at least one group is different.
and at the decision stage, if the result (calculation value) obtained with Equation (3) is greater than Chi-square table value ($\chi^2_T = \chi^2_{1-\alpha;SD}$), absence hypothesis ($H_0$) is rejected; if the calculation value is smaller than or equal to Chi-square table value, absence hypothesis cannot be rejected.

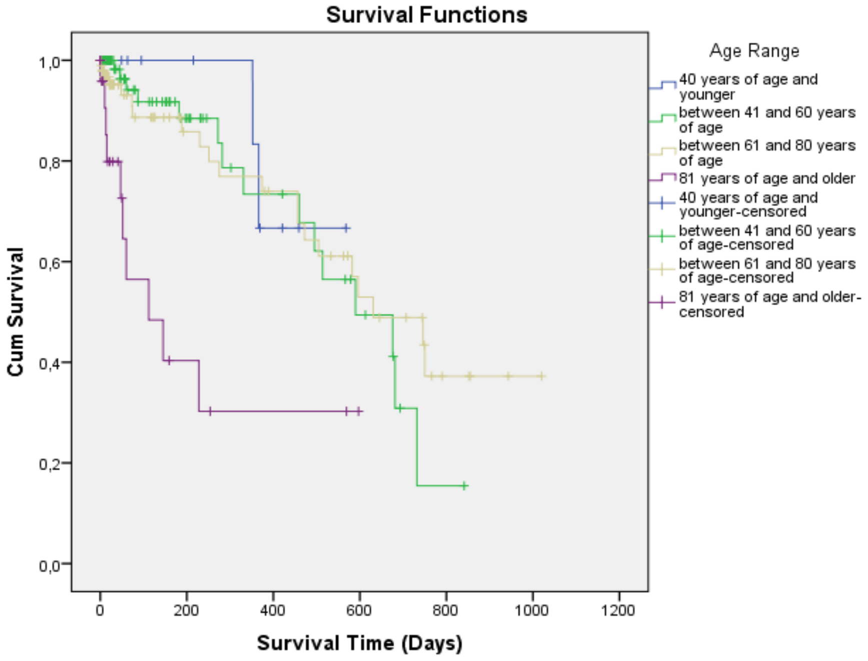
III. MATERIAL

The data of 246 patients diagnosed with colon cancer and rectum cancer were obtained retrospectively from the related unit within the body of Hatay provincial directorate of health. The data including information about the ages, diagnoses, last observation dates and censor states (living, deceased) were taken. The time between diagnosis and last observation dates was calculated; this time was calculated with Kruskal-Wallis test by taking into consideration the state of censor and the results were evaluated. The patients were grouped in four as "40 years of age and younger", "between 41 and 60 years of age", "between 61 and 80 years of age" and "81 years of age and older" so that the calculations could be made.

Kruskal-Wallis test used in the analysis was calculated with the help of MS Excel. While interpreting the results of Kruskal-Wallis test, absence hypothesis was rejected in case of the calculated Chi-square value being greater than the Chi-square table value.

IV. RESULTS

The following results were found as a result of the examinations conducted to find out the associations between four groups: $G^* = \chi^2_{Calculation} = 29.08$ *and* $\chi^2_T = \chi^2_{3;0,05} = 7.81$



In this case, H_0 hypothesis is rejected. In other words, it is said that at least one group is different from the others.

Triple comparisons were made to find out the group that caused the difference. Test results of the first triple comparison groups "40 years of age and younger", "between 41 and 60 years of age", "between 61 and 80 years of age" were found as: $G^* = \chi^2_{Calculation} = 3.00$ *and* $\chi^2_T = \chi^2_{2;0,05} = 5.99$

As can be understood from the results, the calculated Chi-square value is smaller than the Chi-square table value. In other words, absence hypothesis was not rejected and no significant difference was found between "40 years of age and younger", "between 41 and 60 years of age", "between 61 and 80 years of age" groups. Because of this result, no further triple comparisons were thought as necessary.

V. CONCLUSION

In this study which examined the relationship between colorectal cancer and age, the data of a total of 246 colorectal cancer patients who were grouped as “40 years of age and younger”, “between 41 and 60 years of age”, “between 61 and 80 years of age” and “81 years of age and older” were analyzed by using Kruskal-Wallis test, which is a survival analysis method. As a result of the study, a general difference was found in survival times according to age factor when the patients were grouped in four as “40 years of age and younger”, “between 41 and 60 years of age”, “between 61 and 80 years of age” and “81 years of age and older” , when triple comparisons were taken into consideration, the test result was not found to be significant when the last group (81 years of age and older) was excluded from the calculation. In other words, H_0 hypothesis was not rejected and no statistically significant difference was found between the groups in terms of survival time. This shows that the group that creates the difference between four age groups is the “81 years of age and older” group.

The reason why the difference between the survival times of other age groups was not significant can be the fact that a significant part of the data used in the study was censored data and the effects of the disease stages. The evaluation of the stages in the same way was left to further studies.

REFERENCES

[1]. Nelson, W. (1982). *Applied Life Data Analysis*, Kanada: John Wiley & Sons.

[2]. https://fezayarbugkarakayali.com.tr/hastaliklar/kolon-ve-rektum-kanseri/kolorektal-kanser-nedir/

[3]. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RG, Barzi A, et al. Colorectal Cancer Statistics, 2017. CA Cancer J Clin. 2017; 67(3): 177-93

[4]. Özdamar, K. *PASW İle Biyoistatistik*, (8. Baskı. Eskişehir: Kaan Kitabevi, 2010).

[5]. Cox, D. R. and Oakes, D. *Analysis of Survival Data*. London: Chapman and Hall, 1984.

[6]. Bilgi, S. *Sağkalım Analizinde Kullanılan İstatistiksel Yöntemler ve Aktüerya Alanında Bir Uygulama*, Yüksek Lisans Tezi, Ege Üniversitesi, Fen Bilimleri Enstitüsü, 2009.

[7]. Karasoy, D. ve Tilki, B. Yaşam Eğrilerini Karşılaştırmak İçin Kullanılan Skor ve Ağırlıklı Testler: Sayısal Örnekler. *İstatistikçiler Dergisi:* 6, 2013, 1-13.

[8]. G.F. Atkinson and K. Mount, A Nonparametric Test for Type 1 Censored Data: r Samples, *The Canadian Journal of Statistics Vol.22, 1994,149-162.*