ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2023, 24(3), pp. 192-206, DOI:10.18038/estubtda.1335496

RESEARCH ARTICLE

THE USE OF COX REGRESSION MODEL IN THE SURVIVAL ANALYSIS FOR LEUKEMIA PATIENTS IN THE REPUBLIC OF YEMEN.

Elias AL-SAMAI [1], **Sevil ŞENTÜRK** [2]

[1] Department of Statistics, Faculty of Administrative Sciences, Taiz University, Taiz, Yemen
[2] Department of Statistics, Faculty of Science, Eskisehir Technical University, Eskisehir, Türkiye

ABSTRACT

This study aims at analyzing and studying the theoretical and practical importance of the (Cox) regression model in the analysis of survival as well as measuring the most important factors affecting the survival time for patients with leukemia. Moreover, it aims at reaching the expected survival time for patients and creating a life table for patients by using (Cox) regression model. To achieve these goals, real data were taken for (1168) patients with leukemia in the Republic of Yemen in the period from January 2017 to February 2022. The dependent variable, which is the patient's condition at the end of the period, was determined in addition to the patient's survival time and eight independent variables were identified. The effect of these variables on the survival time of patients with leukemia was investigated using the SPSS program. The study concluded several results, the most prominent of them are the following: There are no differences in the incidence rate between males and females and the most age group affected by this disease is (40 years and over). Furthermore, it was found that acute lymphoblastic leukemia (ALL) is the most prevalent type among the other types and there is a difference in the risk of death among those who take intravenous chemotherapy and those patients who take oral chemotherapy. Other significant result was found that there is a higher risk of death for non-regular patients in receiving treatment compared to regular patients in receiving treatment. It was found that the most influencing variables on the survival time of patients are (age, marital status, type of disease, the governorate in which the patient lives, regularity in receiving treatment and type of chemotherapy). Through the life table, it is noticed that the greatest risk in the survival time is in the thirty-sixth month, which is the largest among all other periods, as it reached (0.10). Additionally, the median survival time was reached (35.06) months. Finally, the study found that there are differences in the incidence according to the type of disease in terms of the risk of death as acute lymphoblastic leukemia (ALL) is the most prevalent disease among all diseases and the largest percentage of deaths was among those with chronic myeloid leukemia (CML).

Keywords: Cox Regression, Survival Analysis, Censored, Event

1. INTRODUCTION

Survival analysis is considered as the study of the time of a specific variable until the occurrence of the event as it deals with the time preceding the occurrence of a specific event and one of the most applied examples in this field is the study of the time preceding death. The survival analysis is applied in many different fields such as medicine, engineering, economics, social sciences and others fields where the element of time until the occurrence of a specific event is the main factor for the phenomenon under study.[1,2] In other words, survival analysis is the phrase that is used to describe the analysis of data in the form of times from the original time until the occurrence of a certain event or a certain end point. From the foregoing, it can be said that survival analysis is a set of statistical procedures for analyzing data when the dependent variable (the variable of interest) represents the time until the occurrence of the event. This time may be days, weeks, months or years from the beginning of the item until the

*Corresponding Author:elias.aqlan@taiz.edu.ye
Received:01.07.2023 Published. 22.09.2023

occurrence of the event. The event occurs only once for each item of the study. This event may be death, relapse of the patient...etc. [3,4]

Due to the importance of the topic of survival time and its impact on multiple factors, the urgent need has emerged to develop methods and statistical means to increase accuracy, comprehensive and broad knowledge of the factors affecting the survival of the injured person whether alive or dead within the study period. Among these methods are regression models which are not in their traditional form but rather in a developed form. Such developed form should fit the case of the dependent variable which is bi-response.

One of these models and the most widely used is the (cox) regression model proposed by the English scientist (Cox Dived) in 1972 AD. It is considered one of the appropriate models for binary data through which the survival time and the factors affecting the survival time of the injured person are studied. This model aims at knowing the risk factors that significantly affect the risk function during the duration of time. Furthermore, survival analysis of the Cox regression model involves examining the time from patient admission to the study period until the onset of the event (death) or Censored [5,6].

The problem of the study emerged from the fact that leukemia is one of the most common diseases that leads to the death of thousands of people .Thus, this study focuses on studying and identifying the factors affecting the survival times of leukemia patients in the Republic of Yemen.

In order to identify these factors, Cox regression model was applied as survival analysis is necessary when studying systems in which the dependent variable is the time until a certain event occurs and survival analysis is widely applied in medical and biological studies. Applying such models to different diseases helps in identifying the conditions and characteristics that lead to increasing or decreasing the probability of survival and the factors affecting it. Consequently, the problem of the current study seeks as well at identifying the most important variables affecting the survival time through the Cox regression model and the significant of this study lies on shedding light on survival analysis models especially the (Cox) regression model which makes this study a starting point for other studies.

Mainly, the aims of the current study is based on two main aims which they are measuring the most important influencing factors (risk factors) on the survival time of leukemia patients in the Republic of Yemen by using the (Cox) model and estimating the survival and hazard function of this model as well as forming life table.Consequently the current study identifies the most important factors affecting the survival time of leukemia patients in the Republic of Yemen based on real data form the period January 2017 to the period February 2022 for 1168 patients diagnosed with leukemia.

The hypothesis of the study (H_0) is based on the assumption that there is no significant effect of the variables (sex, age, blood type, marital status, type of disease, the province in which the patient lives, regularity in receiving treatment and type of chemotherapy) on the probability of survival at a significant level of 5%.

2. LITERATURE REVIEW

Okal 2010 studied the survival of breast cancer patients in the Gaza Strip as the study period spanned the period between (2000-2005) ,meanwhile the sample size was 103 women. The Cox model was used by applying (Kaplan-Meier) method to estimate the survival function as the study variables were (date of birth, marital status, address, smoking, date of injury registration, date of end of follow-up, place of appearance of infection, condition, breast containing the primary tissue tumor and treatment). The study concluded that the treatment variables and age are the two influencing variables in the survival time and the rest of the variables had no effect on the survival time.[7]

Meanwhile, the study of S. M. M. Kamal, 2011was based on data of "Social and Economic Determinants of Age at First Marriage among Tribal Women in Bangladesh". This data were collected through a field survey conducted in 2006 in Bangladesh. The study was applied to a sample of 792 currently married women born before 1986. Cox's model was used to study the determinants of age at first marriage for women. The study concluded that women ,who are working in formal jobs, have an impact on women's age at first marriage. The study also indicated that women with higher education are more likely to delay their marriage. The study showed that the place of residence and the educational status of the father have a significant impact on the timing of marriage for women as the study showed that women who were born in rural areas are more likely to marry early. Moreover, women whose fathers are illiterate are more likely to marry early compared to women whose fathers are educated as parents' focus on increasing the educational level of their children. Furthermore, the study indicated that the survival status of the fathers as well as the economic status of the fathers had an impact on the age of the woman at the first marriage. Additionally, the study showed that women whose dowry is less are more likely to marry early. The study indicated also that the order of birth among the sisters has an impact on the woman's age at the first marriage as the sisters take turns in the marriage contract.[8]

Whether, the study Of Burcu Küley Ağir, 2017 aimed at examining survival analysis methods and their application in the field of livestock. In this study, Kaplan-Meier (K-M) and Cox regression methods were applied and the data of two different samples were used in the field of livestock. The first sample is the death records of raising chickens from two different poultry houses and the second sample is the death records of two groups of mice (mice with and without treatment).

In the sample fattening chicken, the data of the rearing period of 5344 chickens were used. The follow-up time for the chickens in the two barns was determined in weeks and both of them were followed up for a period of 23 weeks.

The number of chickens in coop A was 2224 while in coop B there were 3120 chickens. 108 chickens from the first coop experienced the event (death) while 88 chickens from the second coop experienced the event (death). The results showed that there was a significant difference between groups A and B at a significant level (0.01) and the risk of chicken death from the second coop B was 1.74 times higher than that of the first coop A.

With regard to the second sample (mice), it was found that there was a significant effect attributed to the variables of sex and treatment condition and it was found that the time for tumor development in male and female mice was 103,525 and 96,202 weekly respectively. According to the treatment condition variable, the time for tumor development was 100,380 and 98,550 weekly respectively for the control and treatment groups. Finally, it was found that the treatment group had a risk of tumor development 2.193 times higher than the group (without treatment), in addition to that the male mice group had a tumor risk 0.047 times lower than that of female mice [9].

Moreover, S. Selim and S. Sülükçüler's 2023 study aimed at making a comparative analysis of the factors that affect the duration of smoking for individuals in Turkey. The data of health surveys in Turkey for the years 2012 and 2019 were used. The study was conducted on 5932 individuals in 2012 and 6833 individuals in 2019. The study variables were divided into variables (demographic, social, economic, chronic diseases and other variables in four basic categories). The study concluded that the demographic and socio-economic variables, chronic diseases and the reason for starting smoking have a significant effect on the duration of smoking. It was also found that women smoke more than men. In addition, an increase in the level of education and income contribute to an increase in the duration of smoking as well.[10]

Furthermore, Study of Çilengiroğlu Ö. 2023 aimed at finding out the factors that affect months unemployment time for students of the Statistics Department at Eylül University Dokuz 2014-2019 until they can find their first job immediately after graduation using survival analysis methods.

The time that a student takes to find his or her first job after graduation was defined as Event while the time a student takes to find a job as Censored. The results of analyzing the study data by using Kaplan-Meier and Cox regression showed that the variables (gender", higher education after graduation, satisfaction with life, training status and knowledge of computer programs and programming languages are statistically significant for the period between graduation and finding a job. The results also showed that choosing the appropriate place for training is crucial as through appropriate training graduates can find a job faster than expected. It was also found that the average survival of female graduates without work until obtaining the first opportunity reached 6 months while for male students reached 12 months [11] .

3. SURVIVAL ANALYSIS

Survival analysis is defined as a branch of statistics that includes a set of statistical techniques for analyzing data in which the variable of interest is the survival time until the event occurs and time can refer to the age of the individual when the event occurs while the event means the transition from one state to another .In the survival analysis, the event here is death, the occurrence of a certain disease or any particular experience of interest that may happen to the individual,.

As the name of survival analysis is the most widely used and recognized and it relates to the analysis of data that has three main characteristics namely:[12,13]

1. The dependent variable is the residence time until a specific event occurs.
2. The presence of control data.
3. The explanatory variables that affect the survival time and needed to be identified .

The analysis of survival functions includes time modeling for example, the study of the condition of the patient since the diagnosis of injury until the occurrence of the event (the event represents death in the literature of survival analysis in medical experiments) or monitoring (which includes recovery, withdrawal from the hospital without knowing his health condition or death due to a cause other than the reason for the study .Therefore, survival analysis is the only statistical method that deals with controlled and uncontrolled data. [14,15]

- **Survival Data:**

It is an expression used to describe data that measures the time until the event and the resulting variable, which is known as the survival time. This variable is always a positive real variable.

It can be defined as "data that measure the survival time of a group of patients with a specific disease while they are being studied until death or their loss from follow-up due to withdrawal or the end of the study period.

- **The Event**

The concept of the event differs according to the study. In medical studies and research, the event means death from a specific cause (the cause of the study) such as cancer for example. The event can be the emergence of disease, the development of disease, or the relapse of the patient. In industrial applications, the event may be the failure of the unit. In economics, an event may mean getting a job and in demography, it may mean marriage.

- **Censored Data**

The subject of Censored data is one of the topics that have applied importance in the medical and industrial fields and what distinguishes the studies of survival functions or reliability functions from other statistical studies is the phenomenon of monitoring (Censoring) in which part of the information is missing which means that there is partial information about the random variable.

It means units put to the test and the test ends when a certain number of failed units or at a predetermined time. Censoring data often appear in the study of survival especially in medical experiments when the information available about the survival time of the patient under study is incomplete for several reasons, namely:

- The probability that the person does not experience the event before the study ends.
- The possibility that the person will lose himself during the study period.
- The possibility that the person may withdraw from the study because of death (if death is not the event) or for a reason other than the reason for the study.

4. COX REGRESSION MODEL

In the seventies of the last century, various methods and techniques were used to treat the problem of regression in which the dependent variable is subject to Censoring. Some of these methods depend on assumptions about the distribution of survival times while others do not depend on assumptions about the distribution of survival times. One way in which there are no assumptions about the distribution of survival times is the Cox regression model in which it is based on the proportional hazards model .

In 1972, the English scientist (David Cox) estimated the relative risk model or what is named by cox regression model. This model determines the relationship between the explanatory (independent) variables available to the individual studied and the time of survival for them. The cox model is considered the most used model in the analysis of survival data especially when it is used in the case of monitoring data (Censored Survival Time). It aims at knowing the risk factors that contribute significantly to the risk function during the period of time until the emergence of the critical event. In 1975 the English scientist (David Cox) proposed a method to estimate the parameters of the model that he proposed in the year 1972 and he called it the partial possibility function.[16,17]

The model is defined as a statistical method for interpreting the relationship between the patient's survival time and a set of explanatory variables (risk variables) affecting the patient's survival time. The purpose of this model is to explore the effects of a number of variables on the patient's survival. The risk factors for the Cox model are similar to the independent variables in the regular regression models except that they appear in a non-linear exponential form.[18,19]

This model does not assume a specific distribution of survival times rather it assumes that the effect of the various variables is constant over time. It is also called the semi-parametric model because it includes a parametric part. The parametric part is the exponential function of the explanatory variables and a non-parametric part is the baseline hazard function and the model is one of the most important and most common models in survival analysis models as this model is used in cases where the time variable that precedes the occurrence of a particular event is of importance in analyzing the phenomenon in question.

The proposed Cox regression model:[20]

$$\boldsymbol{h(t/x) = h_0(t)\, exp \sum_{i=1}^{p} \beta_i x_i} \tag{1}$$

where is

$\boldsymbol{h(t/x)}$ = represents the conditional hazard function of the model

It is clear from this equation (1) that the conditional risk for a specific individual at time t results from two factors:

$\boldsymbol{h_0(t)}$ **:** The baseline hazard function, which depends on time, expresses the risk function when the independent variables are equal to zero ,unknown, always positive and represents the non-parametric part of the model.

$\boldsymbol{exp \sum_{i=1}^{p} \beta_i x_i}$: It is the relative risk that does not depend on time in which the effect of the independent variables by increasing or decreasing the risk is constant and does not change for the change of time T.

5. DATA ANALYSIS

This section includes the practical application of the cox regression model, the analysis of the most important factors affecting survival times, the clarification of the effect of each variable on the survival time and the preparation of the life schedule. The study data were obtained from the National Cancer Foundation and the National Cancer Control Center in the Republic of Yemen for 1168 patients for the period from January 2017 to February 2022. The most important variables studied were:

- The dependent variable (Status) is the state at the end of the period for the patient which is a descriptive binary variable and indicates the patient's condition at the end of the period.
- **T**: the time variable which is the survival time of the injured person until death or observation and it was calculated on the basis of months.

Eight risk factors have been identified which can be summarized as follows:

1. Patient's gender: male or female.
2. The patient's age was divided into three categories: less than 18 years, from 18 to less than 40 years and from 40 years and over.
3. The patient's blood type: O-, AB+, A+ or O+
4. The patient's marital status: child, single, married, divorced or widowed.
5. The governorate in which the injured person lives and it has been divided into eight categories:

Taiz, Ibb, Aden, Lahj, Abyan, Al-Dhalea, Al-Hodeidah, and others

6. The type of disease which consists of four types:

Acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).
Chronic myeloid leukemia (CML) and chronic lymphocytic leukemia (CLL).

7. The type of chemical treatment used and it was divided into two types:

Intravenous chemotherapy and oral chemotherapy

8. Regularity in receiving treatment has been divided into:

Regular and irregular

Relative Risk Hypothesis Test:
The first step in estimating the cox model is to test the relative risk hypothesis which assumes that the risk rate is constant from one person to another over time within the study.[12]
The relative risk hypothesis is tested in two ways:

1- Drawing Method:
The hypothesis of relative risk is tested by drawing using the (Kaplan-Meier) method where one of the variables is divided into two parts and a curve drawing .Here the variable of the type of chemotherapy was divided into two parts and then a survival curve was drawn for each type as it appears that the two curves are parallel and the difference between them is constant over time. This indicates that the risk rates have a similar behavior for the two sections over the survival time of the infected person and it is noted that the hypothesis of relativity has been achieved as shown in Figure (1).

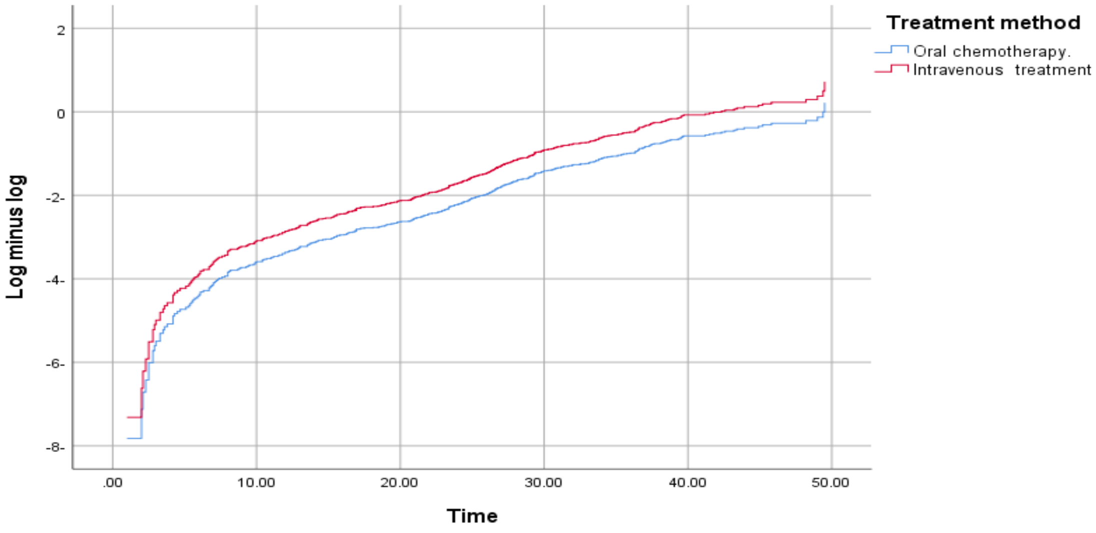

Figure 1. The Graphic Method of Proportional Hazard Hypothesis Test

2- Numerical Met

The imposition of relative risk is confirmed by the numerical method as well by testing (Schoenfeld residuals) errors and according to the null hypothesis which states that the correlation between Schoenfeld errors and survival times is equal to zero as the rejection of this hypothesis means that the relative risk condition is not fulfilled:

$\boldsymbol{H\_0}: \boldsymbol{\delta\_1 = \delta\_2 = \delta\_3 = \delta\_4 = \delta\_5 = \delta\_6 = \delta\_7 = \delta\_8}$**=0**

Table 1. Results of the Schoenfeld Residuals Test to assume Relative Risk

Variables	**Chi-square**	**Df**	**(Sig.)**
Blood type	0.829	1	0.362
Sex	0.022	1	0.882
Marital status	0.593	1	0.441
Region of Living	0.142	1	0.706
Disease type	0.193	1	0.660
Treatment method	0.000	1	0.991
Regularity in receiving treatment	3,630	1	0.057
The age	0.612	1	0.434
Total model	5.209	8	0.735

Through Table (1), it is noted that the calculated value for the overall model that contains 8 variables is (5.209) at (8) degrees of freedom which is less than the tabular value (15.51). This means that it is not significant. Therefore the null hypothesis that states that all correlation coefficients between Schofield errors and survival times are equal to zero is accepted .This indicates that all variables together fulfill the hypothesis of relative risk and the same case for each variable separately. It is found that the value at one degree of freedom for each variable was less than the tabular value of (3.84) and the level of significance for all variables is mor than (0.05). This means that the correlation between Schofield errors and survival times for each variable separately is equal to zero, thus it is identical result to the graph method.

- **Parameter Estimation of Cox Regression Model:**

After realizing the hypothesis of relative risk and estimating the basic functions, the parameters of the cox regression model are estimated by the parametric side of the model using the Partial Likelihood method. It is assumed that the risk function for patients with leukemia is related to the effect of (8) of the variables (sex, age , blood type, marital status, type of disease, the province in which the patient lives, regularity in receiving treatment and type of chemotherapy) on the survival time. Thus, the parameters were estimated using the Partial Likelihood method and Table (2) presents the results of estimating the parameters variables for the (cox) regression model as follows:

Table2. Cox Regression Model Parameters Estimation Results

Variables	B	Wald	SE	df	(Sig.)
Blood Type	0.051	0.401	0.081	1	0.527
Sex	0.016	0.022	0.106	1	0.882
Marital status	0.173	4.115	0.085	1	0.042
Region of Living	0.102	8.363	0.035	1	0.004
The type of disease	0.193	9.404	0.063	1	0.002
Type of treatment	0.453	7.859	0.162	1	0.005
Regularity in receiving treatment	1.127	106.600	0.109	1	0.005
The age	-0.216	4.599	0.101	1	0.032

From table (2), it is noted that the value of significance for the blood type and gender variables amounted respectively (0.527) and (0.882) which means that the coefficients estimated for the two variables are not significant at the level of Significant (5%). Therefore, the null hypothesis, which states that the effect of the two coefficient is zero, is accepted.

Moreover, the value of significance for the (marital status, region of living , the type of disease, type of treatment, regularity in receiving treatment and the age) variables amounted respectively (.042,0.004,0.002, 0.005, 0.005, .032) which means that the coefficients estimated for the variables are significant at the level of significance (5%). Therefore, the alternative hypothesis, that states the coefficients are not equal to zero, is accepted.

- **Testing the Significance of the Variables included in the Model:**

After estimating the parameters of the independent variables of the model, it must be known that the variables which are significant must remain in the model and which are non-significant must be deleted from the model by stepwise deletion of the overall model (Stepwise Backward). The least significant variable is deleted in the form of steps until the significant variables affecting the survival time are reached. This happens through Wald Chi-Squared Test (χ2) ,the calculated (Wald) values are compared with the tabular values at one degree of freedom for each variable and the level of significance is (0.05) as follows:

Table2 :Variables in the Equation

Step	Variables	B	SE	Wald	df	(Sig.)	Exp(B)
Step 1	Blood type	0.051	0.081	0.401	1	0.527	1.052
	Sex	0.016	0.106	0.022	1	0.882	1.016
	Marital status	0.173	0.085	4.115	1	0.042	1.188
	Region of Living	0.102	0.035	8.363	1	0.004	1.107
	The type of disease	0.193	0.063	9.404	1	0.002	1.213
	type of treatment	0.453	0.162	7.859	1	0.005	1.573
	Regularity in Receiving	1.127	0.109	106.600	1	0.000	3.085
	The age	-0.216-	0.101	4.599	1	0.032	0.805
Step 2	Blood type	0.051	0.081	0.400	1	0.527	1.052
	Marital status	0.174	0.085	4.185	1	0.041	1.190
	Living region	0.101	0.035	8.344	1	0.004	1.107
	The type of disease	0.194	0.063	9.461	1	0.002	1.214
	type of treatment	0.455	0.161	7.954	1	0.005	1.576
	receiving regularity	1.126	0.109	106.751	1	0.000	3.082
	The age	-0.217-	0.101	4.651	1	0.031	0.805
Step 3	Marital status	0.171	0.085	4.062	1	0.044	1.187
	Living region	0.102	0.035	8.399	1	0.004	1.107
	The type of disease	0.193	0.063	9.385	1	0.002	1.213
	Type of treatment	0.456	0.162	7.972	1	0.005	1.578
	Regularity in eceiving	1,122	0.109	106.402	1	0.000	3.072
	The age	-0.214-	0.101	4.516	1	0.034	0.807

From table 3 and by comparing the level of significance values (Sig) withe level of significance (0.05), it is noticed in the first step that the variables (marital status, area of residence, type of disease, and type of chemotherapy, regularity in receiving treatment, and age) are the significant variables and that the rest of the variables were non-significant and had no effect on the survival time.

The gender variable had the least effect and it has the lowest significant level of (0.882). Therefore it is deleted from the model in the second step. Moreover, in the second step, it also noted that the variables (marital status, area of residence, type of disease, type of chemotherapy, regularity in receiving treatment and age) are the significant variables except for the blood type variable that has no significant effect as it was less influential on the survival time. The level of significance was (0.527) and therefore it was removed from the model in the third step.

In the third step, it is noted that the survival of the variables (marital status, area of residence, type of disease, type of chemotherapy, regularity in receiving treatment, and age) are all significant variables affect the survival time, thus they remain in the model. The reduced model of the overall model will include the variables that maintained their significance in the last step which are (the marital status, the area of residence, the type of disease, and the type of chemotherapy , regularity in receiving treatment and age). Thus, the shortened form of the (cox) model is as follows:

- **Hazard Function**

$$h(t) = h\_0\ (t)\ \ exp\ (\ 0.171M.d + 0.102K.S + 0.193H.t + 0.456T.m + 1.122T.d - 0.214Yaş)$$

- **Survival Function**

$$\mathrm{S(t)} = [\mathrm{S}_0(\mathrm{t})]^{\mathrm{exp(\ 0.171M.d+0.102K.S+0.193H.t+0.456T.m+1.122T.d-0.214Yaş)}}$$

As for the values of the column (Exp(B)), they represent the estimated risk ratio. If its value is greater than one, it means that the risk is high and if it is less than one, it means that the risk is low. This value expresses the contribution of the independent variable to the risk when it increases more than he correct one in light of the rest of the other independent variables remaining constant.

For example, it is found that the value of (Exp(B)) for the disease type variable is equal to (1.213) which means that the risk rate increases by (21%) for each change in the disease type in light of the rest of the variables remaining constant. It was calculated [(1,213 - 1=-2,13*100=21,3)] are negative sign which means increased risk). To find out whether survival progression differs according to the type of disease, the reference cell coding method was used as the lowest level category which is acute lymphoblastic leukemia (ALL) is selected as a reference cell and compared against the rest of the other categories as follows:

Table 4. Hazard Ration Test for Disease Type Using Reference Cell

The type of disease	B	Wald	(Sig.)	Exp(B)
CLL	1.005	5.789	0.016	2.731
AML	-0.143	0.926	0.336	0.867
CML	0.402	7.367	0.007	1.495

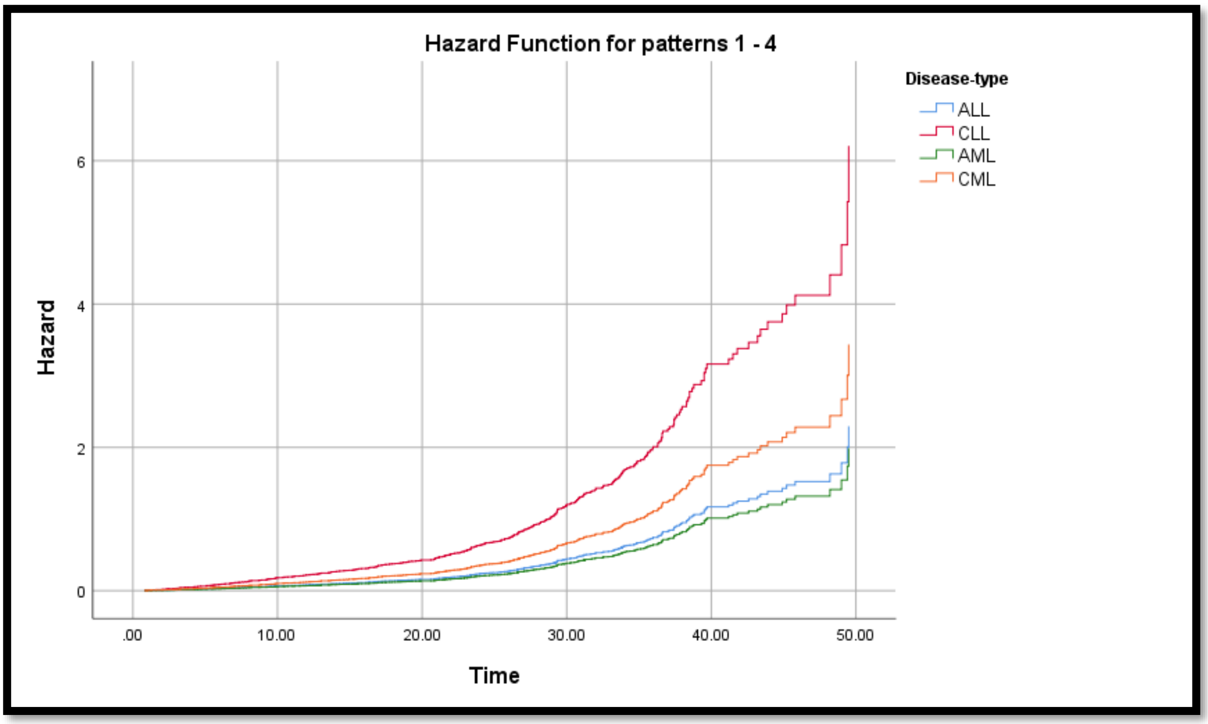

Figure 2. Hazard Ratio for Type of Disease Using the Reference Cell

From table (4) and Figure 2, it is found that the value of the Wald test for type (AML) was non-significant which means there are no significant differences between the risk rate of patients with type (AML) compared with patients with type (ALL. This is due to the level of significance for it was less than (0.05) and the value of the Wald test for patients with both types (CLL) and (CML) was significant.

Therefor, there are significant differences between the risk rate between patients with both types (CLL) and (CML) compared with infected patients with type (ALL). Referring to the Exp(B) value, it is noted that the estimated risk ratio for patients with type (AML) is (0.867) which means that the death rate is lower by (12.4%) for patients with type (AML) than for patients with type (ALL). Moreover, patients with type (CML) die at a rate of (49.5%) more than patients with type (ALL) and it is also noted that patients with type (CLL) die at a rate of (1.731%) times more than patients with type (ALL) .This explains the significance of the variable in the model. Thus, there are at least two of these types that have a risk ratio that differs significantly from the type (ALL).

- **Determine the Best Model:**

When testing the significance of the variables included in the model, several models nominated and the best model in predicting the risk function among all models can be determined through the Likelihood Ratio test which is distributed according to Chi- square distribution as follows:

Table 5: Likelihood Ratio Test Result to choose the Best Model

Step. No _	-2 Log Likelihood	x^2 Calculated	df	(Sig.)
	-2 Log Likelihood = 4395.454			
Step 1	4256.558	138.896	8	0.000 4
Step 2	4256.580	138.874	7	0.000 1
Step 3	4256.972	138.481	6	0.000 0

According to table (5), it is noted that the first model that contains all the variables has a level of significance (0.0004) and that the calculated value was greater than the tabular value. Thus, it is less important than all other models .As well as for the second and third models, it is found that the level of significance for them was less than (0.05) . Consequently, it was found that the third model that contains variables (marital status, city of residence province, type of disease, treatment method used, regularity in treatment and age) is the most significant model among all other models where the level of significance was (0.0000).

- **Survival Times Analysis:**

Depending on the life table equations and dividing the survival times into equal periods of three months and by using the SPSS program, the following results were obtained.

Table 6. Life Table of Survival Data for Patients with Leukemia

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Hazard Rate
0	1168	31	1152.500	11	0.01	0.99	0.99	0.00
3	1126	63	1094.500	26	0.02	0.98	0.97	0.01
6	1037	78	998.000	26	0.03	0.97	0.94	0.01
9	933	56	905.000	22	0.02	0.98	0.92	0.01
12	855	49	830.500	23	0.03	0.97	0.89	0.01
15	783	31	767.500	26	0.03	0.97	0.86	0.01
18	726	67	692.500	21	0.03	0.97	0.84	0.01
21	638	99	588.500	39	0.07	0.93	0.78	0.02
24	500	88	456.000	33	0.07	0.93	0.72	0.03
27	379	75	341.500	43	0.13	0.87	0.63	0.04
30	261	41	240.500	25	0.10	0.90	0.57	0.04
33	195	55	167.500	29	0.17	0.83	0.47	0.06
36	111	32	95.000	25	0.26	0.74	0.35	0.10
39	54	10	49.000	8	0.16	0.84	0.29	0.06
42	36	7	32.500	5	0.15	0.85	0.24	0.06
45	24	11	18.500	2	0.11	0.89	0.22	0.04
48	11	7	7.500	4	0.53	0.47	0.10	0.00

The median survival time is 35.06

Table (6) consists of 9 columns, where each row is the period of stay and each period is divided into three months as the start time of the period starts from zero. It also includes the period up to the third month and the second period starts from the third month to less than six months and so on to 49 months. the second column represents the number of persons entering each period and shows the number of those still alive until the beginning of the current period .It is calculated as $n_{i_{-1}} = n_i - (c_i + D_i)$ where it is found that the number of persons entering at the beginning of the period is 1168 persons while those who remained less than six months in the second period n_2= 1168-(31+11)=1126 .

Their number was 1126 people and so on for the rest of the periods. The third column represents the number of people observed while the fourth column estimates the size of the actual sample which means how many cases were observed during the entry into the time period. For example the number of those entered at the beginning of the second month is calculated as (1126 - (63/2) = 1094.5) and an assumption is made with the case withdraws in the middle of the time period. The fifth column represents the number of whom the event (death) occurred in the time period. For example, at the beginning of the third month, it is found that whom the event (death) occurred were 26 people. The sixth column represents the percentage of the dead while the seventh column is the percentage of the survivors. For example, there are 855 cases who entered the twelfth month.

The probability that the patient completed twelve months is (1-0.03 = 0.97). The eighth column represents the probability of survival at the end of each period***t_i***, for example, the cumulative percentage of patients surviving until the end of the fifteenth month is (0.86) or (87%). The ninth and last column represents the risk function and is known as the probability of failure during a very small period of time. Assuming that the patient survives from the beginning of the period to its end, it was found that the greatest risk is at the time of survival in the thirty-sixth month when it reached (0.10). In addition, it was found that the median survival time is (35.06) months and that the probability of survival at this time is equal to (0.50). This is the time in which half of the cases under study are expected to survive or it is the time after which 50% of the individuals under the study are expected to survive.

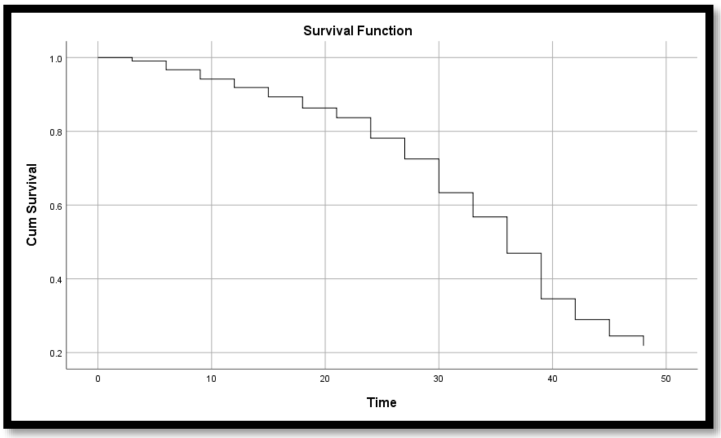

Figure 3. Curve of the Survival Function for Patients with Leukemia according to the Method of Life Tables**.**

Figure (3) shows the curve of the cumulative survival function in the life table and for all periods where each period is equal to three months. The vertical axis represents the cumulative survival function and the horizontal axis represents the time in months . It is noticed that the probability of survival times for infected patients is constantly decreasing until the probability of their survival (21) months which is (0.78) ,the probability that they will stay for a period of 24 months is (0.72) and the probability that they will stay for (45) months is (0.22).

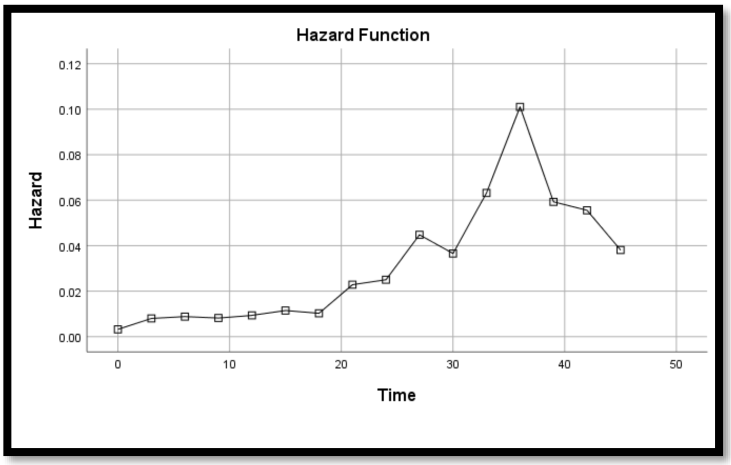

Figure 4. Curve of the Risk Function for Patients with Leukemia according to the Method of Life Tables.

Figure (4) represents the risk function for the life table and for all periods where each period is equal to three months. The vertical axis represents the risk function and the horizontal axis represents the time in months. It is noticed that the risk is low at the beginning of the period then begins to be increased at the twentieth month and then returns to be decreased in the thirtieth month as well as it returns to be raised until the thirty-sixth month then turns to be decreased in the thirty-ninth month. Furthermore, it is found that the highest risk rate is in the period (33-36) months as the risk rate in this period is (0.10) which is the largest rate among all other periods.

6. CONCLUSIONS

This study aimed at shedding light on the theoretical and practical importance of the (Cox) regression model in the analysis of survival, in addition it aimed at measuring the most important factors affecting the survival time of leukemia patients in the Republic of Yemen using the (Cox) model. Moreover, beside the above aims, this study aims at suggesting the best statistical model to determine the degree of risk facing patients with this disease, to estimate the survival and risk functions for this model and to form a Life Table.

All previous studies agreed on the importance of using the Cox regression model in survival analysis and these studies generally aimed at measuring the factors affecting survival time for a group of different phenomena some of which were applied in the medical field as in our study,some of which were applied in the economic field and others in the social field.

In each field, the variables of the study differed according to the study community and the time period spent in applying the study. This study concluded several results, the most important of which are the following:

- The (Cox) model does not depend on a specific distribution of survival times. Consequently, it can be used to indicate the effect of independent variables (risk factors) on the probability of survival as the model provides estimates for the coefficients of each variable. This allows evaluating the effect of multiple variables at the same time and showing their effect on survival time for people with leukemia.

- The results of the analysis have shown that the independent variables (risk factors) that maintained survival in the Cox model are: (age, marital status, type of disease, the province in which the patient lives, regularity in receiving treatment and type of chemotherapy) and it had a significant effect in the time of survival. However, the rest of the variables (sex and blood type) have not shown any significant effect on the time of survival. Furthermore, the model that contains the variables (age, marital status, type of disease, the governorate in which the patient lives, regularity in receiving treatment and chemotherapy type) is the best model among other models.

- The possibility of using the cox regression model in calculating both the survival and risk functions at any given time after estimating the survival and risk functions as well as the basic functions. Furthermore, the possibility of using life tables in analyzing the study data.

- Through the life table, it is noticed that the greatest risk in the survival time is the thirty-sixth month as it reached (0.10) which is greater than all other periods and that the median survival time was (35.06) months.

- There are differences in the incidence according to the type of disease in terms of the risk of death as acute lymphoblastic leukemia (ALL) is the most prevalent disease among all diseases with a rate of (65.8%) of the total patients. The proportions of patients with acute myeloid leukemia (AML) and chronic myeloid leukemia (CML) were close to (16.8%) and (14.7) respectively. Moreover, the highest percentage of deaths was among patients with chronic myeloid leukemia (CML) which reached to (33.1%).

CONFLICT OF INTEREST

The authors stated that there are no conflicts of interest regarding the publication of this article.

AUTHORSHIP CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Elias Al-samai and Sevil Senturk. All authors read and approved the final manuscript.

REFERENCES

[1] Tabachnick BG, Fidell LS, Ullman JB. Using Multivariate Statistics. Pearson Boston, MA; 2007.

[2] Liu ST. SAS, Survival Analysis Techniques for Medical Research; 2004.

[3] Walstra P, Wouters JTM, Geurts TJ. Survival analysis a practical approach. Dairy science & Technology, CRC Taylor & Francis Group. Published online 2005:267.

[4] Lee ET, Wang J. Statistical Methods for Survival Data Analysis. John Wiley & Sons; 2003.

[5] Alrun MB. Survival Analysis of the Registered Colorectal Cancer Cases in the Gaza Strip. BMC Public Health. Published online 2017.

[6] Marshall AW, Olkin I. Life Distributions. Springer; 2007.

[7] Okal M. Survival analysis of breast cancer patients in Gaza Strip. Published online 2010.

[8] Kamal SMM. Socio-economic determinants of age at first marriage of the ethnic tribal women in Bangladesh. Asian Population Studies, 2011;7(1):69-84.

[9] Burcu Küley Ağir. Survival and cox regression analyzes: A Case Study From Animals Science. Published Online 2017.

[10] Selim S, Sülükçüler S. Duration analysis of factors affecting smoking time: A Case Study of Türkiye. Bağımlılık Dergisi. 2023;24(4):475-486.

[11] Çilengiroğlu Öv. Evaluation of the first job finding periods of university graduates with cox regression model. Türkiye Sosyal Araştırmalar Dergisi. 2023;27(1):49-68.

[12] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. Springer Science & Business Media; 2006.

[13] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457-481.

[14] Allison PD. Survival Analysis Using SAS: A Practical Guide. Second Edi. Sas Institute; 2010.

[15] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 1972;34(2):187-202.

[16] O'Quigley J. Proportional Hazards Regression. Springer; 2008.

[17] Klein JP. Handbook of Survival Analysis; 2016.

[18] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin Of Mathematical Biophysics. 1943;5(4):115-133.

[19] Walters SJ. What Is a Cox Model? Citeseer; 2009.

[20] Lawless JF. Statistical Models and Methods for Lifetime Data. John Wiley & Sons; 2011.