Regression Modeling for Competing Risk Analysis with Leukemia at Nanakali Hospital /Erbil

Muhamad Fareed Majeed
muhamed.majeed@univsul.edu.iq
College of Administration and Economics

Asst. Prof. Dr. Samira Muhammad Saleh
samira.muhamad@univsul.edu.iq
University of Sulaymaniyah

نمذجة انحدار للتحليل المخاطرة المتنافسة مع سرطان الدم في مستشفى ناناكالي / أربيل

الباحث .محمد فريد مجيد أ.م.د.سميرة محمد صالح

كلية الإدارة والاقتصاد / جامعة السليمانية

تاريخ استلام البحث ١٥ / ٢٠٢٢/٦ تاريخ قبول النشر ٣ / ٢٠٢٢/٨ تاريخ النشر ٢٧ / ٢٠٢٣/٦

https://doi.org/١٠.٣٤٠٠٩/aujeas.٢٠٢٣.١٨٠٦٣٠

Abstract

The Cox regression model is one of the models that is frequently used in the analysis of survival data, used to determine the relationship between the explanatory variables available for the studied item and their survival time. The aim in this study is to analyze the survival time of patients with leukemia using the two statistical models (Cox regression model and competing risk model). The data used in this study is the Type I of censoring observational data that was taken from Nanakali Hospital-Erbil for (١٢٠) patients with leukemia during (four years) starting from (January ١, ٢٠١٩) to (May ٣٠, ٢٠٢٢). The Akaike Information Standard (AIC), the Corrected Akaike’s Information Criterion (AICc) and the Bayesian Information Standard (BIC) are used for each model to compare two models, which model fits the data. As a result, it shows that the competing risk model fits the cause and that the factor (type of cure) and the (Anemic Condition) factor are the most dangerous for leukemia.

Keywords: *Survival Analysis, competing risk, Cox regression model, Hazard Function, Cumulative Incidence Function (CIF), Leukemia*

المستخلص

نموذج cox للأنحدار هو واحد من النماذج التي يتم استخدامه بالكثرة في تحليل بيانات للبقاء يستخدم لتحديد العلاقة بين المتغيرات التفسيرية المتوافرة للمفردة المدروسة و وقت البقاء لهم. وان الهدف في هذه الدراسة هي تحليل مدة بقاء لمرض المصابين بسرطان الدم (اللوكيميا) بأستخدام النموذجين إحصائين (نموذج انحدار كوكس ونموذج المخاطر المتنافسة). ان البيانات المستخدم في هذه الدراسة هي بيانات المراقبة من نوع الأول والتي تم أخذها من مستشفى نانكلي–اربيل ل(١٢٠) من المرض المصابين بسرطان الدم خلال (أربعة سنة) يبدأ من (١ يناير ٢٠١٩) الى (٣٠ مايو ٢٠٢٢). ويتم استخدام معيار المعلومات اكاكي Akaike (AIC)، معيار معلومات Akaike المصحح (AICc) و معيار المعلومات بيز Bayesian (BIC) لكل نموذج لمقارنة بين نموذجين أي نموذج من النموذجين يلائم البيانات. ومن خلال تحليل البيانات البقاء يبين أن النموذج المخاطر المتنافسة يناسب مع البينات وان العامل (نوع العلاج) والعامل (فقر الدم) هما الأكثر خطورة على مرض اللوكيميا.

الكلمات الأفتتاحية: *تحليل البقاء ، المخاطر المتنافسة ، نموذج انحدار كوكس ، وظيفة الخطر ، دالة الوقوع التراكمي، اللوكيميا .*

١: Introduction

In scientific and biological research, the analysis of event time data or survival statistics aims to describe the hazard (risk) function of event times in populations. In medical research, this event is frequently referred to as "death." Survival analysis is a field of statistics that focuses on dataset analysis with the resultant variable being the time left until an occurrence of interest appears. **(Biost, ٢٠٠٤).**

In survival analysis, we present a test and the model: Kaplan Meier estimation, which is the best statistical method for investigating data in survival analysis, which is used to determine the similarity and differences between two samples, such as treatment and control groups **(Qi, ٢٠٠٩).**

In these cases, the conventional starting point is when the patient enters the hospital, and the endpoint is when the patient dies or lives (censored). An overview of survival analysis is given, as well as significant models pertinent to the current study **(Alhasawi, ٢٠١٥).**

And the exploration and description of non-parametric model: The Cox-PH model, which is currently the most widely used for the investigation of survival analysis in the presence of covariates or prognostic factors, and the Accelerated failure time models, which is a good alternative to the Cox-PH model (Weibull, Exponential, Gamma) **(Wienke, ٢٠١١).**

٢: Literature Review or Related Works

C. Stihsen et al., (٢٠١٧) The objective of this paper is dealing with chondrosarcoma of the pelvis are currently available. Our objective was to identify risk factors for the development of complications. Overall survival was ٧٦٪, ٥٥٪ and ٤٥٪ at one, five and ten years post-operatively. Endo prosthetic reconstruction significantly increases the risk of complications (p = ٠.٠٠٦). Complications were not significantly related to age or the location or grade of the tumor.

Michele Provenzano et al., (٢٠١٨) Hyperkaliemia Under nephrology care, the burden of non-dialysis chronic kidney disease (CKD) is not known. In ٤٦ nephrology clinics, we prospectively monitored ٢٤٤٣ patients over the course of two visits (referral and control with a ١٢-month interval). In the ٣.٦ years of follow-up, ٣٤٩ people died and ٥٦٧ patients developed ESKD. In ٧٩ percent of patients, renin-angiotensin-system inhibitors (RASI) were administered. and had no effect on mortality.

Frank Emmert-Streib and Matthias Dehmer (٢٠١٩) In this study, we looked at the theoretical foundations of survival analysis, such as survival estimators and hazard functions. The Cox Proportional Hazard Model is discussed in length, as well as methods for evaluating the proportional hazard (PH) assumption. We also talk about stratified Cox models for when the PH assumption doesn't hold. Our presentation is supported by a worked example that uses the statistical programming language R to demonstrate how the concept can be applied in practice.

Abderrahim Oulhaj et al., (٢٠٢٠) A slew of COVID-١٩ research that looked into mortality and recovery used the Cox Proportional Hazards (Cox PH) model, which ignores the presence of competing hazards. We study the bias in predicting the hazard ratio (HR) and absolute risk reduction (ARR) of mortality when competing hazards are disregarded, and provide an alternative method based on large simulations. If recovery and mortality due to COVID-١٩ are not included as competing risk events in COVID-١٩ research, there is a significant risk of misleading outcomes. We strongly advise re-analyzing relevant published data that has employed the Cox PH model using a competing risk strategy.

Bsrat Tesfay et al., (٢٠٢١) The objectives of this paper were to find out what factors affected the time to death among breast cancer patients who received anti-cancer treatment at Ayder Comprehensive Specialized Hospital from September ٢٠١٥ to December ٢٠١٨. Methods: Breast cancer patients were studied in a hospital-based retrospective cohort research. The Kaplan-Meier survival curve was used in conjunction with the log-rank test to look for differences in survival across predictor factors.

٣: Methodology (Theatrical Part)

This section studied some basic concepts of survival analysis; survival function, cumulative hazard function and some tests and methods used to analysis survival data, Cumulative incidence function Kaplan Meier, Chi-square test.

٣.١: Survival Analysis (Time-To-Event)

Survival analysis is a branch of statistics that studies how long it will take for an event to occur, such as death in biological organisms or mechanical system failure. In engineering, this is known as reliability theory or reliability analysis; in economics, it is known as duration analysis or duration modeling; and in sociology, it is known as event history analysis. Survival analysis aims to answer specific issues; such as what percentage of a population will live through a certain point in time. **(BALAKRISHNAN & RAO, ٢٠٠٤) (GUO, ٢٠١٠)**

It is required to define "lifetime" in order to respond to such concerns. Death is clear in biological survivability, but failure may be ambiguous in mechanical reliability because there are mechanical systems in which failure is partial, a question of degree, or not otherwise confined in time. Even with biological difficulties, some events (such as a heart attack or other forms of organ failure) might be unclear. **(LIU, ٢٠١٢)**

٣.٢: Survival Function

The survival function is the probability that the survival time, T, is greater than the specific time t; then is characterized as: **(FOX, ٢٠١٤)**

$$\boldsymbol{S}\ (\boldsymbol{t}) = \boldsymbol{Pr}\ (\boldsymbol{T} > \boldsymbol{t})$$

If (T) represents the cumulative time of the life of a particular person during the period (٠, t), and to find the relationship between the probability function and the survival function, we suppose

that the life time (t) of the system is distributed according to the aggregate probability function [$F(t)$], then:

$$\boldsymbol{S}(\boldsymbol{t}) = \mathbf{1} - \boldsymbol{pr}(\boldsymbol{T} \leq \boldsymbol{t})$$

$$\boldsymbol{S}(\boldsymbol{t}) = \mathbf{1} - \boldsymbol{F}(\boldsymbol{t})$$

So that **[$F(t)$]**is sometimes called the improbability function.

٣.٣: Some Basic Definition

٣.٣.١: Censoring

When the time to an occurrence is not recorded for a number of reasons, it is referred to as censoring. There is a lot of filtering in survival analysis. Observations are suppressed when there is insufficient information regarding a subject's time of survival. There are numerous varieties of censorship, including: -

٣.٣.٢: Type I of censoring

The research comes to an end at a certain time or, if the individuals are examined at separate times, when a specified amount of time has passed. **(XIN, ٢٠١١).**

٣.٣.٣: Type II of censoring

When a certain number of incidents have occurred, the research comes to an end **(EKMAN,٢٠١٧)**. We describe this remark as being censored whereas the importance of an observation or measurement is mostly accepted. **(HARRELL, ٢٠٠١)**

٣.٣.٤: Right censoring

If failure happens after the reported follow-up period, a subject is appropriately censored **(MARK, ٢٠٠٧).**

٣.٣.٥: Left censoring

If it is known that the failure occurred before the specified follow-up period, the subject is censored **(HEAGERTY, ٢٠٠٥).**

٣.٣.٦: Interval-censoring

Some of the transition periods have not been observed, although they are known to fall within a certain range. Dementia, for example, has a latent onset period. When longitudinal data is provided, the onset can be determined to be between two consecutive observations. **(HOUT, ٢٠١٧)**

٣.٣.٧: Independent Censoring

Independent censoring has been assumed, which means that after adjusting for covariates, the risk of a censored subject experiencing a future event is the same as the risk of other participants who have the same covariate values and are still being followed up on **(Ekman, ٢٠١٧).**

٤: The cumulative incidence functions (CIF)

The cumulative incidence function (CIF), which is specified independently for each event type and rises over time, formalizes the CIF. The chance that an event of that kind happens at any time point from baseline and time t is known as the CIF at time t. If the data set contains censored observations this obvious estimate must be altered to properly take censoring into account. **(Marcel Wolbers ٢٠٠٩).**

٥: The cause-specific hazard function

The cause-specific hazard function measures the instant potential per unit time for a specific event type to occur at time t among subjects without any prior event. It is calculated as the likelihood per unit period of seeing that event type within a short length of time after time t at time t. We will describe the simplest method for estimating the hazard function, which includes separating the time frame into discrete time intervals. The event of interest happening within a certain interval of time divided by the number observation period during that same timeframe is the definition of the estimated incidence of an event occurrence for that time interval. **(Michael T. Koller ٢٠١٢)**

٦: The Regression Model

As it is known that regression models are those models that study the relationship between the dependent variable and several other variables, which are independent variables and can be expressed mathematically as follows:

$$Y_i = \beta' x_i + e_i$$

Where β are the parameters of the model, which determines the type of relationship determines the type of here between dependent variables as well as the independent variables There are several types of regression models, which depend on the type the explanation's (Y_i) link with the response changes (x_i) on the type of evidence and the problem to be studied, as these differences lead to the difference in the methods of estimating the parameters of the β model, e_i is the Error Term.

٧: Cox Regression Model:

The Cox model of the Cox proportional hazards is the most extensively used and used multivariate method in survival analysis. In the year ١٩٧٢, Cox introduced the Cox model for the first time. There are a variety of Cox homogeneous regression models, but the hazard function is the variable to consider. The hazard ratio is an estimate of the hazard rate ratio based on event rates comparison. **(Ekman ٢٠١٧)**

٧.١: Hazard function

The hazard function is a function denoted by the symbol h(t), and it provides the probability of failure about overall survival, what really is characterized as being the likelihood of a problem caused while a brief presumed to be a number of days person may have remained alive up until the event. Moreover, the independent failure in the brief period of time every unit of time supposing which the person has survived until time (t); **(WIENKE, ٢٠١١) (SCHMIDT & WITTE, ١٩٩٨)**

$$\mathbf{h_{(t)}} = \lim_{\Delta t \to 0} \frac{\mathbf{pr(t \le T \le t + \Delta t | T \ge t)}}{\mathbf{\Delta t}}$$

$$\mathbf{h_{(t)}} = \frac{\mathbf{f\,(t)}}{\mathbf{S(t)}}$$

Let T indicate the duration of an event and have a probability density function. $\boldsymbol{f}\ (\boldsymbol{t})$ and cumulative function $\mathbf{F(t)} = \mathbf{Pr(T \le t)}$, The definition of the survival functional S(t) is: **(FOX, ٢٠١٤)**

$$\mathbf{S(t)} = \mathbf{Pr(T > t)} = \mathbf{1} - \mathbf{F(t)}$$

The hazard function, also known as immediate risk, conditional failure rate, and mortality rate of a specific age, calculates the risk of failure for each unit of time during operation. The data survives since there are no suppressed observations. The risk function is the percentage of patients who die per unit time while knowing they were alive at the start of the period or:

$$\mathbf{h(t)} = \frac{\textbf{number of patients dying per unit time of the interval}}{\textbf{number of patients surviving at t}}$$

٧.٢: The cox proportional hazards regression Model

The most practical method for building regression for survival analysis, time to event, and covariate values is the Cox regression model. The Cox (١٩٧٢) proportional hazards (PH) model had recently gained a lot of popularity as a regression model for the analysis of survival data. These associated survival processes are as follows.: **(BALAKRISHNAN & RAO, ٢٠٠٤)**

$$\boldsymbol{S(t|x) = S_{o(t)}\ exp\ (\sum_{i=1}^{p} B_i X_i\ )}$$

Here, $\boldsymbol{S_{o(t)}}$ is a baseline survival function without variables X that is not stated. The hazard is multiplied by the variables. The exponential and Weibull are obviously exceptions. The risk of one topic is a multiplicative copy of that of another; the model is given as follows when comparing object j to subject m: **(MAWLOOD, ٢٠١٩)**

$$\boldsymbol{\frac{h(t|x_j)}{h(t|x_m)} = \frac{exp(x_j B_x)}{exp(x_m B_x)}}$$

A parametric regression model based on the exponential distribution: (FOX, ٢٠١٤)

$$\boldsymbol{log\, h_{i(t)} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \cdots + \beta_k x_{ik}}$$

Or equivalent:

$$\boldsymbol{h_{i(t)} = exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \cdots + \beta_k x_{ik})}$$
$$\boldsymbol{= e^{\alpha} + e^{\beta_1 x_{i1}} + e^{\beta_2 x_{i2}} + \cdots \cdots + e^{\beta_k x_{ik}}}$$

Where:

$\boldsymbol{i}$ indexes subjects

$\boldsymbol{x_{i1}\ , x_{i2}\ , \dots\ , x_{ik}}$ Are the values of the covariates for the $\boldsymbol{i^{th}}$ subject? Let $\boldsymbol{h(t|x_t)}$ indicate the hazard rate over time for a particular person with a covariate value $\boldsymbol{x_t}$

$$\boldsymbol{h(t|x_t) = h_0(t) * exp(\beta' x)}$$

Here $\boldsymbol{x_t = (x_{1t}\, , x_{2t}\, , \dots\ , x_{kt})\, , \beta = (\beta_1\, , \beta_2\, , \dots\ , \beta_k)}$

$\boldsymbol{k}$**:** is the overall count of covariates.

$\boldsymbol{\beta}_k$**:** Is the treatment's proportional, constant effect.

$\boldsymbol{h}_{0\,(t)}$**:** is known as the baseline hazard function, and it represents the individual's risk when the values of all the independent variables are zero. **(SCHMIDT & WITTE, ١٩٩٨)**

As indicated by **Hosmer and Lemeshow (١٩٩٩)**; in Cox regression the measure that is analogous to $\mathbf{R^2}$ in multiple regression is:

$$\mathbf{R_p^2 = 1 - exp\left[\frac{2}{n}(L_0 - L_p\right]}$$

Where:

$\boldsymbol{L_0}$: is the model's log likelihood with total count parameters.

$\boldsymbol{L_p}$**:** is the log likelihood of the model that includes the covariates.

$\boldsymbol{n}$**:** is the number of observations (censored or not).

The proportionate hazard assumption is an important factor in assessing whether a process's hazards are proportional or not. Some processes need partitioning of failure time, others require categorization of covariates, and still others require the use of a spline function.

٨: Kaplan Meier:

Let time be partitioned into a fixed sequence of intervals.$T_0, T_1, \dots\, , T_k$. These intervals are almost always, but not necessarily, of equal lengths. The survival function of the Kaplan-Meier method is formed as follows:

$$h(t) = \frac{d_i}{n_i}$$

$$s(t) = 1 - h(t)$$

$$s(t+1) = \prod_{k=0}^{t-1} s(k)$$

KM is a non-parametric survival function estimator that is commonly used to define survivability of study participants and to compare dual study populations. In ١٩٥٨, Kaplan Meier (KM) created a cooperative trail and published a paper on dealing with time-to-event data. Later on, KM curves and survival data estimations have proven to be a more effective means of assessing data in cohort studies. **(KOROSTELEVA, ٢٠٠٨)**

The KM estimator of the survival function given as the equation

$$\hat{S}(t_i) = \prod_{t_i \le t} (1 - \frac{d_i}{n_i})$$

Where:

$\mathbf{t_i}$**:** is duration of study at point i

$\mathbf{d_i}$**:** is number of deaths up to point $\mathbf{i}$

$\mathbf{n_i}$: is number of individuals at risk just prior to $\mathbf{t_i}$

S: is based upon the probability that an individual survives at the end of a time interval, on the condition that the individual was present at the start of the time interval, S is the product (P) of these conditional probabilities.

The survival probability S(t) is a periodic function, and only an event causes its value to change. the confidence levels for survival probabilities can indeed be easily determined by calculating. The survival curve plots the KM survival rate as a function of time, and gives a beneficial overview of the information may prepared to calculate metrics that is average survival time.

٩: The Log Rank Test

The Log Rank test is a non–parametric method for testing the null hypothesis that groups are samples of the same survival experience. It is applicable to data where there is progressive censoring and gives equal weight to early and late failures (**Vittinghoff, ٢٠٠٤**).

١٠: Competing Risk Analysis

Competing risk analysis is a sort of survival analysis that aims to predict the marginal probability of an event in the presence of competing occurrences properly. Normal forms to describe survival processes aren't built to take into consideration the competing nature of multiple reasons for the same event. To address this issue, the Cumulative Incidence Function (CIF) was proposed, which calculates the probabilities of various occurrences as a result of their overall survival likelihood and cause-specific survival likelihood. Competing-risks analysis is a variation on traditional survival analysis. When there are competing events to a primary endpoint, risk estimates will be biased. **Zhou, Bingqin (٢٠١١).**

١١: Model Selection Criteria:

١١.١: Chi-Squared Test:

It is determined using the Chi-Squared test whether or not a model came from a group having a certain distribution. The test statistic's value is decided by way data is discarded because it is a test performed on binned data. Keep in mind that this test is only applicable to continuous sample data. The Chi-Squared statistic has the following definition:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i}$$

Where:

c**:** is the degrees of freedom

O_i**:** is your observed value

E_i: is your expected value.

It’s very rare that you’ll want to actually *use* this formula to find a critical chi-square value by hand. The summation symbol means that you’ll have to perform a calculation for every single data item in your data set.

١١.٢: Akaike’s Information Criterion (AIC)

Utilizing the Akaike Information Criterion (AIC) assesses it level and excellence of a collection of statistical methods. Akaike's Information Criterion, for instance, is determined Following is: **(Moore ٢٠١٦)**

$$AIC = -2loglikelihood + 2K$$

Where:

K: is how many model parameters there are (the intercept means the number of constructs in the model). A measurement model fitting called log-likelihood is typically acquired based on statistical results.

١١.٣: The Corrected Akaike’s Information Criterion (AICc)

There is a high likelihood that AIC will choose models with too many parameters when the sample size is small, or that it will overfit. AICc was created to address this potential fitting problem: AICc stands for AIC with a small sample size correction. The statistical model affects the AICc formula. The formula for AICc is as follows if the model is univariate, linear in its parameters, and has normally-distributed covariances**. (Burnham, Kenneth P. and David R. Anderson ١٩٩٨)**

$$AICc = AIC + \frac{2k^2 + 2k}{N - k - 1}$$

AIC: Akaike’s Information Criterion

K: No. Of Parameter

N: Sample Size

١١.٤: Bayesian Information Criterion (BIC)

One of the most well-known and often used tools for choosing statistical models is the Bayesian information criterion (BIC). Each model's BIC is calculated, and the model with the lowest BIC value is chosen.

$$BIC = -2loglikelihood + 2 * logN * k$$

L is the likelihood value, N is the total amount of observations, while k is the total amount of calculated parameters.

Calculating each model's BIC is all that is necessary to compare them using the Bayesian information criteria; the model with the least BIC is deemed to be the best model. **(Lee, T. & Wang, W. ٢٠٠٣).**

٤: Application Part

٤.١: Data Collection & Description:

The Kurdistan Regional Government of Iraq, the Ministry of Health, the General Administration of Health, and the Nanakli Hospital, which specializes in hematology in the Erbil Governorate, provided the information for this study on leukemia. The data included ١٢٠ cases that were gathered during a ٤-year period, starting on ١ January ٢٠١٩ and ending on ٣٠ May ٢٠٢٢, of any and all leukemia patients who were monitored by the hospital through ٣٠ May ٢٠٢٢. Out of those patients, ٢٣ patients passed away during the trial, ٥٣ patients survived or are still living, while ٤٤ competing events occurred. Statistical methods were employed to examine the data (STATA, SPSS, Easy fit).

Description of Data:

Table (٤-١) below show the all variables that used in this Work

Factors	Classification	N	# Of Death	# Of Alive	# Of Risk
Age	١=١-١٠	٦	٣	١	٢
	٢=١١-٢٠	١٠	٥	١	٤
	٣=٢١-٣٠	١١	٣	١	٧
	٤=٣١-٤٠	٢٤	٨	٥	١١
	٥=٤١-٥٠	١٦	٦	١	٩
	٦=٥١-٦٠	١٩	٧	٦	٦
	٧=٦١-٧٠	٢٢	٨	٢	١٢
	٨=٧١-٨٠	١٠	٣	٤	٣
	٩=٨١-٩٠	١	١	٠	٠
	١٠=٩١-١٠٠	١	٠	٠	١
Gender	٠= Male	١٠٢	٦٧	٣٥	٤٧
	١= Female	١٨	٨	١٠	٨
Type of Disease	٠= Acute myeloid leukemia (AML)	٤٦	٢٣	٨	١٥
	١= Acute lymphocytic leukemia (ALL)	١٤	٦	٠	٨
	٢= Chronic lymphocytic leukemia (CLL)	٣٤	١٢	٧	١٥
	٣= Chronic myelogenous leukemia (CML)	٢٦	٣	٦	١٧
Type of Cure	٠= Chemotherapy	٧٩	٢٤	٠	٥٥
	١= Biotherapy	٤١	٢٠	٢١	٠
Address of Patients	٠= Outside the City	٩٩	٣٩	١٥	٤٥
	١= Inside the City	٢١	٥	٦	١٠
Anemic Condition	٠= Infected	٩٥	٣١	١٩	٤٥
	١= Not Infected	٢٥	١٣	٢	١٠
Status	٠= Death	٤٤	…	…	…
	١= Alive (Censored)	٢١	…	…	…
	٢= Competing Event (Risk)	٥٥	…	…	…

٤.٢Testing and statistical Analysis of Data:

٤.٢.١: Hypothesis testing

At the first in order to know the shape of the distribution of survival times for patients with leukemia, this is done by drawing the histogram of survival times as in Table (١), where we find it difficult to know the shape of the real distribution. We use the chi-square test (χ^2) of good fit (χ^2, goodness of fit) test for some known distributions, However, to no use this test

H_o*: *The data ~ specific distribution

H_1*: *The data ~ specific distribution

Table (٢): The results of the good-matched chi-square test to determine the type of distribution, survival times

form distribution	Calculated chi-square value	Chi-Square tabular value of	Degrees of freedom	P-value
Weibull	١٦.٨٤	١٥.١	٥	٠.٠٠
Exponential	١٧.٤٣	١٥.١	٥	٠.٠٠
Gamma	١٧.٧٣	١٥.١	٥	٠.٠٠

Table (٢) shows the results of the chi-square test for good conformance to the models that were tested for the test through equation, as it turns out that the data are not subject to any of these known distributions. From the above table we note the p-value of all the distribution are less than (α=٠.٠١) and we accept the alternative hypothesis that the data are not specific distribution.

٤.٣: Kaplan Meier Test

In order to describe and contrast the survival of two research groups, Kaplan Meier (KM), a non-parametric survival function estimate, is frequently used. The most often used summary statistics in survival analysis are the mean and median. The mean admission time, on the other hand, allows us to predict how many days a patient will need to live with a specific admissions incidence without knowing the whole mean time to event.

Table (٣) The Means and Medians for Survival Time for (Type of Cure) in each group

Mean and Median for Survival Time								
Mean					Median			
Type of Cure	Estimate	St. Error	Confidence Interval		Estimate	St. Error	Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
٠ (Chemotherapy)	٦٣٧.٧	٥٣.٤	٥٣٣.١	٧٤٢.٣	٨٤٤.٠	٤٤.٣	٧٥٧.١	٩٣٠.٩
١ (Biotherapy)	٥٠٤.٨	٦٧.٤	٣٧٢.٨	٦٣٦.٨	٥٢٤.٠	١٤٤.٣	٢٤١.٢	٨٠٦.٨
Overall	٥٨٦.٧	٤٢.٦	٥٠٣.١	٦٧٠.٣	٧٨٦.٠	١٥٧.٩	٤٧٦.٣	١٠٩٥.٧

Table (٣) gives the results of KM test for Type of Cure factor applied to a data set of size ١٢٠ cases. This table shows that the Estimated mean time for patients who Receiving Chemotherapy Cure, is ٦٣٧.٧ days while who does not receive chemotherapy cure (Biotherapy) is ٥٠٤.٨ days with the confidence interval (٥٣٣.١, ٧٤٢.٣) for receiving Chemotherapy cure and (٣٧٢.٨, ٦٣٦.٨) for who Receiving Biotherapy cure under probability ٩٩٪. However, patients receiving Biotherapy have better chance of survival than those receiving Chemotherapy, However, to survive, the patient must receive chemotherapy if the effects of the disease have increased.

Mean and Median for Survival Time								
Mean					Median			
Gender	Estimate	St. Error	Confidence Interval		Estimate	St. Error	Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
٠ (Male)	٦٠٣.٧	٤٧.٧	٥١٠.٧	٦٩٦.٦	٧٨٦.٠	١٨٠.٨	٤٣١.٦	١١٤٠.٤
١ (Female)	٥١٦.٧	٩٣.٩	٣٣٢.٤	٧٠٠.٩	٨٠٥.٠	٤٣٦.٦	٠.٠٠٠	١٦٦٠.٧

Overall	٥٨٦.٧	٤٢.٧	٥٠٣.١	٦٧٠.٣	٧٨٦.٠	١٥٧.٩	٤٧٦.٣٢٣	١٠٩٥.٧

Table (٤) The Means and Medians for Survival Time for (Gender) in each group

Table (٤) explains the estimated mean time until death for males is ٦٠٣.٧ days. While for females is ٥١٦.٧ days with the confidence interval (٥١٠.٧, ٦٩٦.٦) for male and (٣٣٢.٤, ٧٠٠.٩) for female under probability ٩٩٪. In contrast, the median estimated time between Leukemia cancer and death for both male and female. Female patients are more likely infected and have a less chance of survive than the Male.

Table (٥) The Means and Medians for Survival Time for (Anemic Condition) in each group

Mean and Median for Survival Time								
Mean					Median			
Anemic Condition	Estimate	St. Error	Confidence Interval		Estimate	St. Error	Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
٠ (Yes)	٣٠٥.٨	٧٤.٨	١٥٩.٢	٤٥٢.٤	١٨٥.٠	١٣.٤	١٥٨.٨	٢١١.٢٠٤
١ (No)	٦٤٦.٧	٤٥.٧	٥٥٧.٢	٧٣٦.٢	٨٤٤.٠	٤٣.٢	٧٥٩.٣	٩٢٨.٧
Overall	٥٨٦.٧	٤٢.٧	٥٠٣.١	٦٧٠.٣	٧٨٦.	١٥٧.٩	٤٧٦.٣	١٠٩٥.٧

Table (٥) displays the estimated mean time for patients, who infected anemia is ٦٤٦.٧ days while who does not infected is ٣٠٥.٨ days with the confidence interval (٥٥٧.٢, ٧٣٦.٢) for infected anemia and (١٥٩.٢, ٤٥٢.٤) for does not take radiotherapy under probability ٩٩٪. Anemia is one of the most important factors in leukemia that affects survival or life expectancy. It is clear from the analysis that patients who are not anemic are more likely to survive.

٤.٤: First Model: Cox Proportional Hazard Model

The Cox regression model (Cox, ١٩٧٢) is essentially a regression analysis commonly medical research uses. The Cox-PH model is a well-recognized statistical technique for exploring the relationship between patient's survival and several explanatory variables. After accounting for other explanatory variables, the Cox PH model gives an estimate of the impact of the treatment on survival. In six variables, the model-building process takes place (Age, Gender, Type of Disease, Type of Cure, Address, Anemic Condition). In our study, there are ٧٥ censored cases (which represent the number of patients who are still alive but have not yet died) and ٤٥ event events (which represent the number of deaths). Additionally, the case is also said to be censored if the occurrence has not yet happened.

Table (٦): Results of fitting a Cox Proportional Hazard model

Factors	β	Std. error	Wald	Degrees of Freedom	P-value	Exp (β)	Confidence Interval	
							Lower Bound	Upper Bound
Age	-٠.٠٤٩	٠.٠٧٨	٠.٣٩٨	١	٠.٥٢٩	٠.٩٥٢	٠.٨٢	١.١١
Gender	٠.٧٥	٠.٣٨٢	٣.٨٤٤	١	٠.٠٥٠	٢.١١٦	١.٠٠	٤.٤٨
Type of Disease	-٠.٥٣٠	٠.١٤٨	١٢.٨١٢	١	٠.٠٠٠	٠.٥٨٩	٠.٤٤	٠.٧٩
Type of Cure	٠.٤٨٥	٠.٣١٧	٢.٣٥٢	١	٠.١٢٥	١.٦٢٥	٠.٨٧	٣.٠٢
Address of Patients	-١.٥٨٠	٠.٥٠٣	٩.٨٥٤	١	٠.٠٠٢	٠.٢٠٦	٠.٠٧	٠.٥٥
Anemic Condition	١.٢٣٨	٠.٣٧٦	١٠.٨٣٢	١	٠.٠٠١	٣.٤٤٩	١.٦٥	٧.٢١

Table (٦) is the result or Cox regression analysis. For this analysis, we will discuss the values separately and explain them briefly. By this way (Exp (β), St. Error, Wald, Significances, And the Upper Bound or Lower Bound of Confidence Interval). However, we explained the confidence interval with Exp (β) value.

- Exp (-٠.٠٤٩) = ٠.٩٥٢, which is a lowering in the risk of mortality for patients with age (x_1), indicates that age is one of the factors influencing the risk of leukemia malignancy. The considerable value indicates that the leukemia cancer is significantly impacted.
- One of the factors influencing the risk of dying from leukemia and other cancer disorders is gender (x_2). Exp (٠.٧٥) =٢.١١٦, which represents an increase in the likelihood that a patient with (Male or Female). The p-value is ٠.٠٥٠, indicating that any gender is more likely to die from cancer.٢
- • After accounting for the other explanatory factors in the patient's death model, the estimated risk for the Type of Disease (x_3) is Exp (-٠.٥٣٠) = ٠. ٥٨٩, a ٥٨.٩ percent drop in the risk. Also, the ٠.٠٠٠ p-value indicates statistical significance.
- Explanatory factors in the death model for chemotherapy patients Type of Cure (x_4). However, the hazard ratio's ٩٩ percent confidence interval is included, the p-value of ٠.٠٣١ is statistically significant, and the estimation for risk is increased by Exp (٠.٤٨٥) = ١.٦٢٥ for the Type of Cure component.
- The estimate hazard in the address of the patient (x_5) group is, Exp (-١.٥٨) = ٠.٢٠٦ which is ٢٠.٦٪ After adjusting for either explanatory variable, the chance of death for all patients decreased.
- Anemic condition (x_6) is among the elements that influences risk of leukemia cancer. For increase by Exp (١.٢٣٨) = ٣.٤٥ which is increase in the likelihood that a patient having (Male or Female). Given that either gender has a higher likelihood of dying after cancer, the p-value is ٠.٠٠١.

When all of the fixed covariates' vectors ($\underline{x}$) are present (Gender, Type of treatment, Type of Disease, Address of Patient and Anemic Condition) while β is the equivalent vector for the fixed covariates' regression model.

$$(t)= \beta_0(t) * exp(\beta_i{}'x)Y_i$$

Y_i(t)= β_0(t) * exp (٠.٧٥ Gender-٠.٥٣ Type of disease+٠.٤٨٥ Type of treatment-١.٥٨٠ Address of patient+١.٢٣٨ Anemic condition

Because of the rating statistics with number of powerful or more ٠.٠١, which is two variables, the aforementioned model does not accept these variables. (Gender and Type of treatment), has not significant.

In the table above if the value of Wald column is considered as a significant factor; then, Type of disease will be one of the significant factors in our study; because it has a greater value in Wald test column (١٢.٨١٢) with significant value of (٠.٠٠٠<=٠.٠٠١).

Table (٧) Covariate means for all factors

Covariates Means	
Factors	**Means**
Age	٥.٠٠٨
Gender	٠.١٥١
Type of Disease	١.٣٤٥
Type of Cure	٠.٣٣٦
Address of Patients	٠.١٧٦
Anemic Condition	٠.٢١٠

Table (٧) displays the average value of each predictor variable; this table is a useful reference for survival plots, which are constructed for the mean values and each pattern.

٤.٥: Kaplan-Meier Curve

For each target group, a crucial component of the survival analysis is the graph of KM curves. The curve of the hazard function with the cumulative hazard is the most significant curve in our study. The graph of the cumulative hazard ratios for the various treatment groups (i.e., the two kinds of queries is the equivalent vector for the fixed covariates' regression model. Status), "Dead" and "Alive", is typically used to interpret our results.

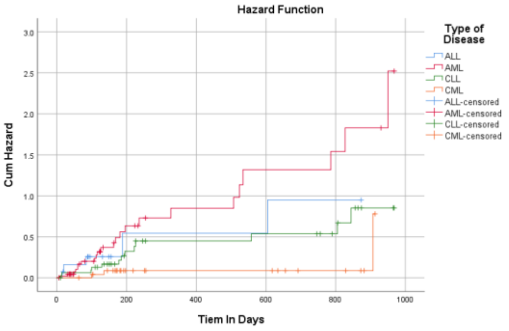

Figure (١): Kaplan-Meier graph Type of Disease

Figure (١) shows the cumulative hazard on the time and the x - axis to incident on the y-axis. The plot of the hazard curve makes it obvious that the danger of dying rises with time, sometimes stabilizing before rising once more. We can easily observe that the risk of passing away has not decreased. in particular for the AML.

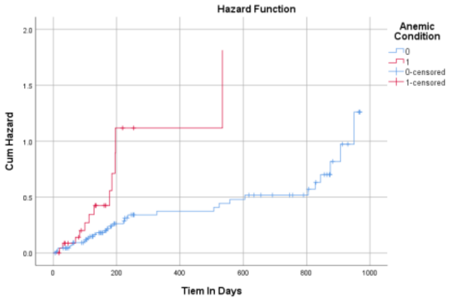

Figure (٢) Kaplan-Meier curve of Anemic Condition

The danger of dying increases over time and occasionally stabilizes before increasing once more in Figure (٢), where the horizontal axis shows the time to event, and the axis the cumulative hazard. We can easily observe that the risk of passing away has not decreased.

Table (٨) the Results of Log Rank test (Anemic Condition)

Overall Comparison			
	Chi-Square	Degrees of Freedom	P-value
Log-Rank (Mantel-Cox)	١٢.٢١٣	٣	٠.٠٠٠
Test of equality of survival distributions for the different levels of Type Disease			

Table (٧) explains that the p-value is ٠.٠٠٠≤٠.٠١ which indicates that there is a significant difference between the two groups (infected with Anemia and not infected with Anemia) on having a short time to event. The estimated time until death is ٦٤٦.٦٧ days for patient infect with Anemia, ٦٢٦.٣٢ and ٣٠٥.٧٩ for patient who doesn't infect with Anemia.

Table (٩) the Results of Log Rank test (Type of Disease)

Overall Comparison			
	Chi-Square	Degrees of Freedom	Sig.
Log-Rank (Mantel-Cox)	١٥.٢٥٠	٣	٠.٠٠٢
Test of equality of survival distributions for the different levels of Type Disease			

Table (٨) explains that the p-value is ٠.٠٠٢≤٠.٠١ which indicates that there is a significant difference between the two groups (infected with Anemia and not infected with Anemia) on having a short time to event. The estimated time until death is ٤٠٥.٣٤ days for patient diagnosis with (AML) Type of Disease, ٥٠١.٠٧ days for patient who diagnosis with (ALL) Type of Disease, ٦٢٦.٣٤ days for patient diagnosis with (CLL), and ٨٤٣.٧٨ for patient who diagnosis with (CML) Type of Disease.

Table (١٠): Second Model: First Model Competing Risk Regression

Factors	Sub Hazard Ratio	Std. error	Z	P-value	Confidence Interval	
					Lower Bound	Upper Bound
Age	٠.٩٩	٠.٠٠٧	-٠.٨٧	٠.٣٨٣	٠.٩٩	١.٠٠٨
Gender	٢.٣٢	٠.٧٣٧	٢.٦٥	٠.٠٠٨	١.٢٤	٤.٣٢
Type of Disease	٠.٦٥	٠.٠٨٨	-٣.١٢	٠.٠٠٢	٠.٥٠	٠.٨٥
Type of Cure	٣.٧٣	١.١٤٣	٤.٢٩	٠.٠٠٠	٢.٠٤	٦.٧٩
Address of Patients	٠.٣٦	٠.١١٣	-٣.٢٦	٠.٠٠١	٠.١٩	٠.٦٧
Anemic Condition	٢.٠٨	٠.٦٩٤	٢.٢١	٠.٠٢٧	١.٠٩	٤.٠٠٣

We have estimated the Sub-Hazard Ratio in table (١٠) by a series of ways (lowest risk, medium risk and highest risk with confidence interval for each variable) and then we have discussed the variables that are significant, and the last thing to note is that the variables that are significant are significant and their effect on the disease.

١. The highest risk in the data we analyzed was for variable (Type of Cure) with a ratio of (٣.٧٣) and (٢.٠٤ to ٦.٧٩), Another risk in the data we analyzed was for variable (Gender) with a ratio of (٢.٣٢) and (١.٢٤, ٤.٣٢).

٢. The medium risk in the data we analyzed was for variable (Anemic Condition) with a ratio of (٢.٠٨) and (١.٠٩, ٤.٠٠٣), A second medium risk comes to (Age) variable with a ratio of (٠.٩٩) and (٠.٩٩, ١.٠٠٨).

٣. The last and The Lowest risk in the data we analyzed was for variable (Type of Disease) with a ratio of (٠.٦٥) and (٠٥٠ to ٠.٨٥), and other last and lowest risk in the data we analyzed was for variable (Address of Patient) with a ratio of (٠.٣٦) and (٠.١٩, ٠.٦٧).

That sub hazard ratios are greater than one to be expected, given the non-parametric estimates of the cause-specific (CIF) having higher values for the higher risk groups.

While the significances for those factor that distributed with this data is (Gender significant by (٠.٠٠٨), Type of Cure significant by (٠.٠٠٠), Address of Patients significant with (٠.٠٠١), Type of Disease (٠.٠٠٢), and the last significances Value comes to Anemic Condition with (٠.٠٠١)).

Table (١١): Fitting the model using Akaike's Information Criterion (AIC), Akaike's Information Criterion Corrected (AICc) & Bayesian Information Criterion (BIC)

Model	No. of observation	Log Likelihood	AIC	AICc	BIC
Cox Regression	١٢٠	-١٥٧.٠٣٢	٣٢٦.٠٦٤	٣٢٦.٨٠٧	٣٤٢.٧٩
Competing Risk Regression	١٢٠	-٨١.٨٢	١٧٥.٦٤	١٧٦.٣٨٣	١٩٢.٣٧

The results are shown in Table (١١) for the AIC and BIC values that were used to make comparisons between different models (Cox regression model and Competing risk regression) in order to choose the model that best fits our data on leukemia cancer. The findings indicate that the competing risk regression model is better for our leukemia research data since its AIC, AICc, BIC values are the lowest when compared to the Competing Risk Model AIC and BIC values of ١٧٥.٦٤, ١٧٦.٣٨٣ and ١٩٢.٣٧, respectively.

٤.٦: Cumulative Incidence Function (CIF) Curve

With each group of interest, a key component of Competing Risk is the plot of CIF curves. The cumulative incidence curve is the most significant curve in our analysis. The graph of the cumulative incidence functions for the various treatment groups (i.e., the two types of queries regarding Status), "Alive" and "Death" is typically used to interpret our data.

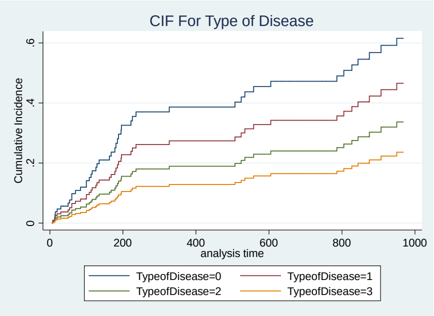

Figure (٣) CIF curve for Type of Disease

The cumulative hazard is represented by the vertical axis in Figure (٣) and the time to event is represented by the horizontal axis. It is obvious from the plot of the hazard curve that the danger of dying rises with time, sometimes stabilizing before rising once more. It is obvious that the risk of dying is not diminishing. In particular, for the Type of Disease=٠ (AML). According to the graph, disease is increasing

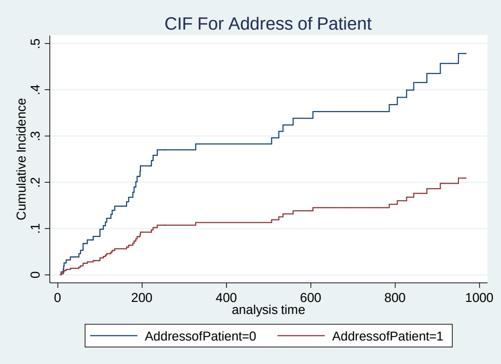

Figure (٤) CIF curve for Address of Patient

The risk of leukemia in the graph over time outside the city is gradually decreasing after the availability of hospitals and necessary health services outside the city.

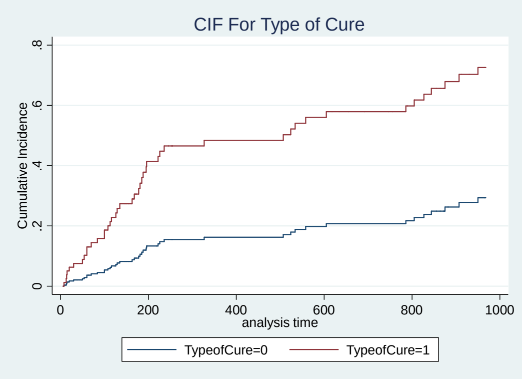

Figure (٥) CIF curve for Type of Cure

Over time, chemotherapy is gradually decreasing due to the emergence or discovery of a specific cure for cancer in general.

Table (١٢) Comparing Hazard Ratio (Cox Proportional Hazard Model) & Sub-Hazard Ratio (Competing Risk Regression)

Cox Proportional Model	Competing Risk Regression
Anemic Condition= ٣.٤٤٩	Type of Cure= ٣.٧٣

To compare the hazard ratio between the two models and to determine the highest risk in each model, so the (Cox Proportional Hazard Model) variable has the highest risk of all variables with a (Anemic Condition) variable value of (٣.٤٤٩) and the (Competing Risk Model) variable has the highest risk of all variables There are more variables with a (Type of Cure) value of (٣.٧٣) It is obvious that people with leukemia need treatment (٠= Chemotherapy, ١= Biotherapy) to reduce the risk to life.

Competing Risk Regression & Cox (Table (١٣) Combined value between two models Regression)

Cox Proportional Model	Competing Risk Regression
Type of Disease= ٠.٠٠٠	Type of Disease= ٠.٠٠٢
Address of Patients= ٠.٠٠٢	Address of Patients= ٠.٠٠١

After separating the variables that share the same significance, the competing risk regression model gives the sum of both factors as ٠.٠٠٣, and the Cox proportional hazard model gives the sum of both factors as ٠.٠٠٢, finally Both factors are significant, but the Cox proportional hazard model is closer to significance.

Conclusion

From the results obtained from the applied chapter, we can infer that the most important conclusions reached by the study are:

١. The Cox-PH model does not identify the same prognostic factors that influence in leukemia survival time.

٢. The results of cox regression model of this study illustrated that the most important factors that effecting on the leukemia in using data set are (Type of Disease and Anemic Condition).

٣. Comparing the results of the Competing Risk regression model with the Cox regression model based on the AIC, BIC and AICc criterions it is concluded that Competing Risk regression model has smallest value of AIC, BIC and AICc criterions or is the most suitable model for the data set used in this study.

٤.

Recommendations

١. To improve the Cox regression model and the Competing Risk regression model, We could improve the number of attributable factors, such as some pertinent risk factors and family histories of leukemia patients, that are reliable predictors of survival time. These would aid in understanding the traits of health-related behaviors connected to leukemia patients' survival rates.

٢. We hope that the results of this thesis will be taken into consideration by ministry of health and Nanakali Hospital-Erbil.

٣. There will be communication between the Ministry of Health and researchers so that the following researches will be more expensive, powerful and more significant.

٤. Another suggestion for data collection for scientific research is to facilitate access to medical data in health centers or hospitals, because it is very difficult to provide data to higher education students.

Reference

١. Abderrahim Oulhaj et al., ٢٠٢٠.The competing risk between in-hospital mortality and recovery: A pitfall in COVID-١٩ survival analysis research. Institute of Public Health, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates.

٢. Alhasawi, E., ٢٠١٥. Survival Analysis Approaches for Prostate Cancer. Sudbury, Ontario, Canada: Laurentian University.

٣. Balakrishnan, N. & RAO, C. R., ٢٠٠٤. Handbook of Statistics ٢٣_ Advances in Survival Analysis. north holand: s.n.

٤. Biost, A. (٢٠٠٤). Introduction to Survival Analysis. new yourk: A Stata Press Publication.

٥. Bsrat Tesfay et al., ٢٠٢١. Survival analysis of Time to Death of Breast Cancer Patients: in case of Ayder Comprehensive Specialized Hospital Tigray, Ethiopia. Corresponding author: Endeshaw Assefa Derso, College of Natural and Computational Science, Department of Statistics, University of Gondar, Ethiopia.

٦. Burnham, Kenneth P. and David R. Anderson. ١٩٩٨. Model Selection and Inference: A Practical Information-Theoretical Approach. New York: Springer-Verlag.

٧. C. Stihsen et al., ٢٠١٧. The outcome of the surgical treatment of pelvic chondrosarcomas: a competing risk analysis of ٥٨ tumors from a single center, Medical University of Vienna, Vienna, Austria.

٨. Ekman, A., ٢٠١٧. Variable selection for the Cox proportional hazards model. Umea university, ٢١ January.p. ٨٤.

٩. Fox, John, ٢٠١٤. Introduction to Survival analysis. sociology ٧٦١.

١٠. Frank Emmert-Streib and Matthias Dehmer. ٢٠١٩. Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, FI-٣٣١٠١ Tampere, Finland

١١. Guo, S., ٢٠١٠. Survival analysis. ١st ed. New York: Oxford university press, Inc.

١٢. Harrell, F. E., ٢٠٠١. Regression Modeling with Applications to Linear Models, Logistic Regression, and Survival Analysis. Nashville, TN ٣٧٢٣٢-٢٦٣٧: Springer Science+Business Media New York.

١٣. Heagerty, P., ٢٠٠٥. Survival Analysis. new work: Va/Uw Summer.

١٤. Hout, A. V. D., ٢٠١٧. multi-state survival models for interval censored data. U.S.: taylor & francis group.

١٥. Mark, s., ٢٠٠٧. An Introduction to Survival Analysis. EpiCentre, IVABS, pp. ٢-٣١.

١٦. Liu, x., ٢٠١٢. Survival analysis 'models and applications'. ١st ed. United Kingdom: John Wiley & Sons Ltd.

١٧. Michele Provenzano et al., ٢٠١٨. Competing-Risk Analysis of Death and End Stage Kidney Disease by Hyperkaliemia Status in Non-Dialysis Chronic Kidney Disease Patients Receiving Stable Nephrology Care. Journal of Clinical Medicine.

١٨. Qi, J. (٢٠٠٩). Comparison of Proportional Hazards and Accelerated Failure Time Models. Saskatchewan, ١-٨٩.

١٩. Schmidt, P. & WITTE, A. D., ١٩٩٨. Predicting Recidivism Using Survival Models. ١st ed. London: Springer verlag.

٢٠. Vittinghoff, E. ٢٠٠٤. Statistics for Biology Health, Second edition. Science+Business Media, New York.

٢١. Wolbers, marcel, et al., ٢٠٠٩. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology ٢٠.٤ (٥٥٥-٥٦١).

٢٢. Wienke, A., ٢٠١١. Frailty models in survival analysis. ١st ed. USA: Taylor & Francis group, LLC.

٢٣. Xin, x., ٢٠١١. A STUDY OF TIES AND TIME-VARYING COVARIATES IN COX PROPORTIONAL HAZARDS MODEL. University of Guelph, pp. ١-٤٣.

٢٤. Zhou, Bingquing et al., ٢٠١١. “Competing risks regression for stratified data”. Biometrics (٦٦١-٦٧٠).