Fitting cox proportional model and Poisson regression model to data of patients with stomach cancer in Erbil-Kurdistan/Iraq

Kurdistan Ibrahim Mawlood [1], Chnar Smko Abdullah[2]

[1, 2] Salahaddin University, Erbil College of Administration and Economics - Statistics & Informatics Department

***Abstract*— This study aims to fitting two models where allow the response variable to be the length of time (months) to data of patients with stomach cancer; cox proportional model and Poisson regression model, for modeling and identifying the affecting factors of stomach cancer patients. The study was conducted between January 1, 2016 until December 31, 2020 for all patients with stomach cancer at Nanakali Main Hospital for Cancer in the Kurdistan Region of Iraq - Erbil.**

The results indicated that, the models have not reached to the same variables that have an impact on our data of patients with stomach cancer data in Erbil city. Moreover, according to the results the Poisson regression fitted data set very well depending on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, the best model will be identified by the ones with smaller values. The data analyses of stomach cancer are done by using statistical programs (Mat-lab V.14 , SPSS V. 25 and STATGRAPHICS V. 19).

***Keywords*— Survival Analysis, cox proportional model, Poisson regression model, Akaike Information Criterion (AIC), stomach cancer.**

I. INTRODUCTION

Survival analysis is the process of analyzing data based on time-to-event (survival times). The time to event data shows the period of time between a well-defined time origin and a well-defined end point of interest (event). Although time-to-event analysis and time-to-event data are often used more frequently than survival analysis and survival data, the term "Time to event" is clearer and more accurate. It's important to clearly identify the time origin and end point. As an illustration, in a study of a specific type of cancer, the diagnosis of that type of cancer is chosen as the time origin, and the death caused on by that specific cancer is chosen as the time end point. Or a research might examine people from the time of their birth (time of origin) until the appearance of a disease (end point). The data on the time to occurrence is typically collected sequentially over time, such as when data was collected for a clinical experiment or a proposed cohort study. Retrospective data collection methods include speaking with patients who have the disease in question or obtaining access to their medical records.(ISRA, 2019)

1. Background Information

in this section Stomach cancer which is the first important health issue discussed. Two functions, the survival function and the hazard rate also known as the hazard function, are used to describe the stomach cancer data. The survival function measures the possibility that a patient will survive to time t while the hazard rate (function), measures the possibility that the patient will die in the future instant of time. Moreover, exploration, description and the basic principles of two models (Cox regression models and Poisson regression models) given and the Log rank test to compares survival of two different groups of individuals. Also, to select the best model between two models (Cox and Poisson regression model) Akaike's Information Criterion and Bayesian information criterion were used.

2.1 Cancer

Cancer is a chronic disease that begins in the cells that are body's building blocks. Human body functions in a way that when old cells die, new cells are formed to replace the old ones. However, there are conditions where there is a mutation and the process goes wrong and cell formation does not take place in a normal way. It may happen like new cells start forming even without their need in the human body, that can be malignant or benign. If we discuss the benign tumors, they are not considered as cancer, whereas the malignant tumors are definitely very risky as they are base of cancer disease. Cells that are present in the malignant tumors have the possibility of being shifted to other parts of the human body causing serious risk to health condition which is known as metastasis. (AHMAD, 2019)

2.1.2 Stomach cancer

Stomach cancer is portrayed by a development of cancerous cells inside the lining of the stomach. Likewise called gastric cancer, this sort of cancer is hard to analyze in light of the fact that many people ordinarily do not show signs and symptoms in the early stages. While compared to other types of cancer, stomach cancer is often uncommon, one of the biggest risks of this disease is the difficulty in diagnosing it. Since stomach cancer typically doesn't have any early symptoms, it

Polytechnic Journal of Humanities and Social Sciences
Volume 4 No. 2 (2023)
Received 30 july 2022; Accepted 4 September 2022; Regular research paper: Published 15 August 2023

frequently goes undetected until it has spread to other parts of the body, this makes it very hard to cure.

2.1.3 Development of stomach cancer

Usually stomach cancers grow slowly over many years. Pre - cancerous alterations frequently appear in the mucosa of the stomach before a real cancer develops. Since these early changes rarely have symptoms, they may go unnoticed. Cancer cells must undergo a number of modifications before they may move to new areas of the body. In order to adhere to the exterior wall of a lymph conduit or blood vessel, they first need to develop the ability to separate from the primary tumor. Cancerous cells can spread throughout the bloodstream and end up in distant organs. Cancer cells could end up in lymph nodes if they move through the lymphatic system. The majority of the cancer cells that escape either die or are eliminated before they may begin to grow somewhere. However, a few could move, start to spread, and develop new tumors. Metastasis is the medical term for the spread of cancer to a new area of the body. (AMERICAN CANCER SOCITY, 2022)

2.2 Survival Analysis

The two fundamental parts of a survival analysis are the event time and the event status, both of which contain information on the occurrence of the relevant event. The survival and hazard functions, which both depend on time, can be fitted into two categories using event time. For the survival analysis to characterize the distribution for event times, these two functions are essential notions. The survival function provides the probability of surviving up to each individual time point. The probability that the event will happen, per unit of time, is provided by the hazard function. (AMERI, 2015)

The majority of medical researches focus on the event of time to death. However, another crucial factor in cancer is the amount of time that passes between a treatment response and a recurrence or period of disease-free time. Additionally, it's vital to indicate the situation and duration of the observation, such as the period between a cancer diagnosis of confirmed response and the first relapse. The time to event data may include information on patient characteristics related to response, survival, and disease development, as well as information on survival time and treatment response. (ISRA, 2019)

Let T represent an individual's survival time with probability distribution function f and the cumulative distribution function $F(x) = \int_0^x f(u)du$

The survival function, S(t), gives the probability that a subject will survive until time t: (AMERI, 2015)

$S(t) = S(t) = p_r(T \geq t) = 1 - F(t) = \int f(x)dx$ …. (2.1)

In contrast, we can express the probability distribution function as:

$$f(t) = \frac{\partial F_{(t)}}{dt} = -\frac{\partial S_{(t)}}{dt} \quad \text{…. (2.2)}$$

The hazard function serves as the foundation of a survival analysis for a number of reasons. It first tells us whether and, if so, when events occur, which is exactly what we want to know. The risk of the event occurring in each time period is summarized by its magnitude. second the hazard function includes both censored and uncensored situations. Third, no data is ignored or pooled; the sample hazard probabilities are calculated during the entirety of an event's occurrence. Fourth, In time periods where censoring prevents its direct computation, the sample hazard function can be utilized to estimate the sample survivor functions indirectly.

suppose the survival time T is such that $t \leq T$, $t + \delta t$, then this probability can be expressed as:

$P(t \leq T < t + \delta t | T \geq t)$

By dividing by the interval length δt and by evaluating the limit of this conditional probability at δt approaches zero, we obtain a rate which defines the hazard function.

That is, the Hazard is given by:

$$\boldsymbol{h_{(t)} = \lim_{\Delta t \to 0}[\frac{pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}]}$$

$$\boldsymbol{h_{(t)} = \lim_{\Delta t \to 0}[\frac{[F(t + \Delta t) - F(t)]|\Delta t)}{S(t)}]}$$

$$\boldsymbol{h_{(t)} = \frac{\partial F(t)|\partial t}{s(t)}}$$

$$\boldsymbol{h_{(t)}} = \frac{f(t)}{s(t)} \quad \text{…. (2.3)}$$

The Hazard function is also known as conditional failure rate. The failure Hazard per unit time throughout the operation is provided by the Hazard function, which is fundamental to the data. However, in practice, the hazard function is the proportion of patients who die per unit of time when there is no controlled observation, even when they have survived to the starting point of the period.: (HOUT, 2017)

$$h(t) = \frac{\text{number of patients dying per unit time of the interval}}{\text{number of patients surviving at t}} \quad \text{…. (2.4)}$$

2.3 Censoring

In survival analysis, If the occurrence of interest hasn't been observed for a certain person, their survival time is called censored. This might be as a consequence of the fact that the survival data is analyzed since some people remain alive. It could also be due to the fact that some individuals have voluntarily left the experimental or clinical trial without notice. A survival time could also be considered as censored if event of interest death for example was due to a cause that is known to be unrelated to the treatment. Suppose an individual has been recruited into an experimental study at an initial time t_0 dies at time t_0 +t with t unknown due to the fact that the individual is still alive or due to not following-up. If the individual was last known to be alive at time t_0 + c, c > 0, then the time c is called a censored survival time.

2.4 Cox proportional hazard model

An example of an event history model is the Cox proportional hazard model. It regards time as continuous and makes no assumptions regarding the hazard function's shape. Due to the semiparametric Cox proportional hazard model's rising popularity, it's essential to develop practical methods for

determining whether the model is properly described. (KLEINBUM & KLEIN, 2012) provided a graphical procedure, a goodness-of-fit testing strategy, and a procedure involving the use of time-dependent variables as three methods for assessing the proportional hazard (PH) assumption of the Cox model. Numerous articles had discussed the proportional hazard model's graphical and goodness-of-fit methods. (ALJAS, 1988), (MOREAU, et al., 1985) (PARZEN & LIPSITZ, 1999), (WEI, 1984) .

Cox proportional hazard model assumes that the effects of covariates remain constant over time. The discrete-time survival model is more flexible than the Cox proportional hazard model in that it can take into account the effects of covariates that change over time. The Cox proportional hazard model is used frequently in a wide range of fields, yet it has some major limitations. The fundamental presumption that the interaction is cancelled when time is not included in the equation is the most important. declaring that time is essential for time-varying predictors and that time should be taken into account in the model. (SINGER & WILLETT , 1991)

Therefore, the Cox proportional hazard model's survival probability function can be written as follows:

$S(t|x) = S_{o(t)} \exp^{(\beta x)}$. …. (2.5)

Or

$S(t|x) = S_{o(t)} \exp(\sum_{i=1}^{p} B_i X_i)$. …. (2.6)

When;

$s_0(t) = \bar{e}^{\int_0^t h_0(x)\,dx}$. …. (2.7)

The hazard ratio (HR) is constant over time about any two sets of variables,x and x^*

$$\frac{h(t|x^*)}{h(t|x)} = \frac{h_0(t)\,exp\left(\sum_{k=1}^{p}\beta_k x^*\right)}{h_0(t)\,exp\left(\sum_{k=1}^{p}\beta_k x\right)} = exp\,(\sum_{k=1}^{p}\beta_k(x^*\_\,x)\ldots. (2.8)$$

Let $h(t|x_t)$ denote the hazard rate at time for an individual have covariate value x_t

$h(t|x_t) = h_0(t) * \exp(\beta' x)$ … (2-9)

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

k : is the total number of the covariates.

β_k : Is the treatment's consistent proportional effect.

$h_{0\,(t)}$ The baseline hazard function, represents an individual's hazard when all independent variable values are equivalent to zero. (SCHMIDT & WITTE, 1998)

As indicated by Hosmer and Lemeshow (1999); in Cox regression the measure that is analogous to R^2 in multiple regression is:

$R_p^2 = 1 - \exp\left[\frac{2}{n}(L_0 - L_p\right]$ … (2-10)

Where:

L_0 : is the log likelihood of the model with no covariates.

L_p : is the log likelihood of the model that includes the covariates.

n : is the number of observations (censored or not).

Where L_p denotes the log partial likelihood for a fitted model with p covariates, and L_0 represents the log partial likelihood of model zero, a model without any covariates. The partial likelihood ratio test, denoted by the letter G, is calculated as two times the difference between the log partial likelihood of the model with the covariate and the log partial likelihood for the model without the covariate.

$G = 2[\, L_p(\hat{\beta}) - L_p(0)]$ …. (2.11)

Where $L_p(0) = -\sum_{i=1}^{m} ln(n_i)$ and the term n_i represents the number of individuals in the risk group at the observational survival time t_i.

The statistic will follow a Chi-square distribution with p degrees of freedom under the null hypothesis that its coefficient is equivalent to zero.

2.4.1 The assumption of proportional hazards

Here, some key assumptions can be made.

l. First of those assumptions is that the proportional hazard, in a given study, needs to be fixed from one patient to another.

2. The second assumption is that there needs to be a linear relationship between the natural log of the hazard function and the explanatory variables. Along with these two assumptions.

3. The third assumption is that the explanatory variable, in any case, does not need to depend on time.

4, Another key assumption that can be made is that statistical distributions should not be distributed by any response variable involved in the study.

5. Finally, another assumption is that the rate of hazard needs to increase in a linear pattern with time. (COLLECT, 2003)

2.4.2 Estimating the coefficients in the Cox PH model

The baseline hazard function h_0 (t) and the regression coefficients β_1,..., m are the unknowns of the Cox model. An alternative formula for the proportional hazards model. It avoids the knowledge of the functional form of h_0(t). The most commonly used method for estimation of the regression coefficients is the partial - likelihood estimation method. With this approach, the time-dependent factor of a likelihood function is omitted, and the remaining elements maximized, known as the partial - likelihood function, to produce the maximum partial - likelihood estimates of the regression coefficients (β_1,... , β_m).

Probability of (individual i dies at $t_{(j)}$ given one death from the risk set R($t_{(j)}$) at time $t_{(j)}$)

$$= \frac{P\left(\text{individual i dies at } t_{(j)}\right)}{P\left(\text{one death at } t_{(j)}\right)}$$

$$= \frac{P\left(\text{individual i dies at } t_{(j)}\right)}{\sum_{k\in R(t_{(j)})} \text{P(individual k dies at } t_{(j)})}$$

where R (t_j) denotes the set of all subjects who are at risk at time t_j. Replacing the probability of death at time t_j, with the probability of death in the interval [t_j, t_j+ Δ) and passing to the limit as Δ → 0, yields the following expressions: (KOROSTELEVA, n.d.)

$$\cong \frac{\frac{P\left(\text{individual i dies at } t_{(j)}, t_{(j)} + \Delta t\right)}{\Delta t}}{\sum_{k\in R\left(t_{(j)}\right)} \text{P(individual k dies at } t_{(j)}, t_{(j)} + \Delta t)/\Delta t}$$

$$= \frac{h_0\left(t_j\right)\exp(B' x_i t_j)}{\sum_{k\in R\left(t_j\right)} h_0\left(t_j\right)\exp(B' x_k t_j)}$$

$$= \frac{\exp(B\prime x_i t_j)}{\sum_{k\in R\left(t_j\right)} \exp(B\prime x_k t_j)}.$$ …. (2.12)

Where k : represent the number of distinct observed event times, $x_i(t_{(j)})$: is a x covariate vector values of individual i that dies at time $t_{(j)}$ and R_i has been the risk set that contains individuals for whom observation event or censoring time is higher than or equal to T_i (LOKESHMARAN A, 2013) (فتيحة، 2015).

When applied to the Cox PH model, the partial likelihood function is as follows:

$$L_{(\beta)} = \prod_{i=1}^{k} L_i = \prod_{i=1}^{k} \frac{\exp(\beta' x_i t_j)}{\sum_{k \in R(t_{(j)})} \exp(\beta' x_k t_j)}$$

$$L_{(\beta)} = \prod_{i=1}^{k} \frac{\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{\sum_{i=1}^{k} e^{(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}}$$

$$l(\beta) = \log(L_{(B)}) = \log\left[\prod_{i=1}^{k} \frac{\exp(\beta' x_i t_j)}{\sum_{k \in R(t_{(j)})} \exp(\beta' x_k t_j)}\right]$$

$$l(\beta) = \sum_{i=1}^{k} [\beta' x_i - \log\{\sum_{k \in R} exp\ \beta' x_i\}$$

$$U(\beta) = \frac{\partial}{\partial \beta}\, l(\beta) = x_i - \frac{\sum_{k \in R} x_i exp\, \beta' x_i}{\sum_{k \in R} exp\, \beta' x_i}$$

$$\beta' = x_i - \frac{\sum_{k \in R} x_i exp \beta' x_i}{\sum_{k \in R} exp \beta' x_i}$$

$$I(\beta) = -\left[\frac{\partial}{\partial \beta}\left(U(\beta)\right)\right] = -\frac{\partial}{\partial \beta}\left[x_i - \frac{\sum_{k \in R} x_i exp \beta' x_i}{\sum_{k \in R} exp \beta' x_i}\right]$$

$$I(\beta) = -\left[\frac{\sum_{k \in R} x_i x_i' exp \beta' x_i}{\sum_{k \in R} exp \beta' x_i} - \frac{[\sum_{k \in R} x_i exp \beta' x_i]\,[\sum_{k \in R} x_i' exp \beta' x_i]}{(\sum_{k \in R} exp \beta' x_i)^2}\right]$$

…. (2.13)

And then the maximum partial likelihood estimators. Asymptotically have:

$\hat{\beta} \sim N(B_0,\ I^{-1}(\hat{B}))$

Where: $I^{-1}(\hat{\beta})$ represent the information matrix inverse for $\beta = \hat{\beta}$, and $\beta 0$ represent a true value of β. To construct confidence intervals and test the hypothesis, this approximate distribution is utilized. H0: $\beta = \beta_0$, (CAMERON & TRIVEDI, 2012).

2.5 Poisson Regression Model

For data including counts, the Poisson regression model is the most used. any observed count, y_i , is selected from a Poisson distribution in the Poisson regression model, and the mean μi is represented as a vector of predictors X_i to evry i^{th} subject. The probability density function for the Poisson distribution is as follows: (CAMERON & TRIVEDI, 2012).

$$P(Y = y) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad y = 0, 1, 2, \ldots$$

…. (2.14)

Over the range of potential values 0, 1, 2,..., the Poisson distribution is unimodal and right-skewing. The term "equidispersion" refers to the fact that it has a single parameter, μ >0 , which acts for both its mean as well as its variance. (AGRESTI, 2006)

Although the "Poisson regression model" is appropriate for modeling "count data," in reality, typically the variance of count data exceeds its mean, leading to Over-dispersion. The Poisson regression model's count data causes bias in the findings and underestimates the parameters, therefore affects the standard errors and P-value. The unobserved randomized variations element in the function of X' could be the cause of this over-dispersion. (CONSUL & FAMOYE, 1992)

the Poisson regression model is expressed as:

$$E(y_i) = \mu_i$$

…. (2.15)

$$g(\mu_i) = X_i^T \beta$$

…. (2.16)

$$\mu_i = g^{-1}(X_i^T \beta)$$

…. (2.17)

Substituting the log link function gives:

$$ln\mu_i = X_i^T \beta$$

…. (2.18)

$$\mu_i = e^{X_i^T \beta}$$

…. (2.19)

The βs for the regression must be estimated using the maximum-likelihood estimation (MLE) method. Starting with the formulation for the likelihood of observing y as a function of β (where y_i, ..., $y_n$are considered to be independent), the MLE for the Poisson regression can be calculated. (Montgomery, et al., 2006)

the likelihood function therefore is:

$$L(y, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

…. (2.20)

The log likelihood function is:

$$l = ln \prod_{i=1}^{n} \left(\frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}\right)$$

…. (2.21)

$$= \sum_{i=1}^{N} ln\left(\frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}\right)$$

$$= \sum_{i=1}^{N} (-\mu_i + y \ln \mu_i - Lny!)$$

$$= -\sum_{i=1}^{N} \mu_i + \ln \mu_i \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} Lny_i!$$

$$= -n\mu_i + (\sum_{i=1}^{N} y_i)\ \ln \mu_i - \sum_{i=1}^{N} Lny_i!$$

$$= \frac{\partial l}{\partial \mu_i} = -n + (\sum_{i=1}^{N} y) \frac{1}{\mu_i} = 0$$

$$\mu_i = \frac{1}{N} \sum_{i=1}^{N} y$$

…. (2.22)

2.5.1 Assumptions of Poisson Regression Model

In order to verify that the data you wish to analyze can really be analyzed using Poisson regression, must first determine whether you can use Poisson regression to analyze your data. This is necessary because Poisson regression should only be used if your data "passes" the five required assumptions that Poisson regression has to make in order to get a reliable result. It is important to do this, nevertheless, as it happens frequently that data will violate one or more of these assumptions:

1. Count data are used to create the dependent variable. Data obtained for other common types of regression (linear regression, multiple regression, logistic regression) is differs from data obtained for counts. Count variables, on the contrary side, need integer value which must be zero or higher. However, since count data must be "positive" (i.e., comprise "nonnegative" integer numbers), it is impossible for it to contain negative values. Additionally, it is frequently

recommended that Poisson regression use only in cases that the mean count is a low value (e.g., less than 10). A different kind of regression may be more suited when there are a large number of counts.

2. There may be one or more independent variables that are measured on such a continuous, ordinal, nominal, or dichotomous scale. Ordinal and nominal (dichotomous) variables are categorized as categorical variables in general.

3. the observations should be independent of one another. This implies that each observation is independent of the others, i.e., no observation may inform the other observations in any way. This is an essential assumption. Lack of independent observations is mostly a problem with survey methodology.

4. - Based on the model, the counts (the dependent variable) distributed according to a Poisson distribution.

5. The models mean and variance are the same. This is an outcome of Assumption 4, which holds that a Poisson distribution exists. The variance for a Poisson distribution is equal to the mean. there is equidispersion if this assumption is valid. This is not always the case, nevertheless, and your data are either under- or over-dispersed, with over dispersion being the more typical problem.

2.6 Differences between Cox regression models Poisson regression models:

1. The dependent variable: In the Poisson regression, the dependent (Y) variable is an observed count, while in Cox regression the dependent variable (descriptive binary + time until event occurs).

2. Censoring: The Poisson regression model does not deal with censored data while Cox regression model deals with it.

3. Method of estimation of regression coefficients: Poisson regression coefficients are estimated using the Maximum Likelihood Estimation method, while Cox regression coefficients are estimated using the Partial Likelihood Estimation.

4. Time: Cox regression models time to event, and Poisson regression models counts or rates of events.

5. If you chop the time axis into finer and finer pieces, then the model will be equivalent to a cox-regression, and in that case the difference is only that the parameter of the time-effect is non-parametric in the cox-regression while it will be estimated together with other parameters in the Poisson regression model.

2.6 Goodness-of-Fit

After the estimating process is complete, attention may go to evaluating how closely the model fits the data. (Coxe, et al., 2009)

One of the fundamental criteria's of goodness of fit is Pearson's statistic. The equation below gives the Pearson statistics for a model with mean λi and variance ωi are given in Equation below:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\omega}_i} \quad \ldots. (2.23)$$

This statistic is applied to verify if the series' dispersion is over. It will be $\omega i = \lambda i$ as a logical consequence of the Poisson distribution if Pearson statistic is used in Poisson regression, the formula takes the form in the Equation below.

$$p_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad \ldots. (2.24)$$

When the calculated χ_p^2 to degree of freedom ratio is more than 1, it means that the data are too dispersed and are not appropriate for the model.

Deviance statistics is one method for evaluating how well a model fits the data. and the deviance, or G, statistic expressed by the following equation:

$$D_p = \sum_{i=1}^{n} \{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\} \quad \ldots. (2.25)$$

When this statistical value approaches zero, it means that the model is fitting the data in a better way. When the statistical value is zero, the model fit is optimal. The residual deviance needs to be as minimal as possible. The ideal residual deviance for Poisson regression is the number of observations minus the number of parameters, or the remaining degrees of freedom of the model. (SHINGLETON, 2012)

$$R^2 = \frac{LL_{fit} - LL_0}{LL_{max} - LL_0} \quad \ldots. (2.26)$$

Note that LL_0 is the log-likelihood of the intercept-only model, LL_{fit} is the log-likelihood of the current model, and LL_{max} is the maximum log-likelihood possible. The maximum log-likelihood occurs when the actual responses (the y_i 's) exactly equal the predicted responses (the μ_i 's). Notice that this value of R-squared varies between zero and one. (Anon., 2016)

2.7 Measures of the Model Selection

In this study, two measures were used to compare Poisson regression and Cox regression models and to determine which model was the best. The measures for each model were calculated, and thus the model with smallest score was selected to be best model for fitting the data.

2.7.1 Akaike's Information Criterion (AIC) For Selecting Best Model

In order to choose the suitable model, models are compared using Akaike's information criteria (Akaike, 1973). The chosen model is the one with the greatest Kullback-Leibler distance between itself and truth. Therefore, the AIC is defined as:

$$\mathrm{AIC} = -2\mathrm{L} + 2\mathrm{K} \quad \ldots. (2.27)$$

Where k : represents parameters number in the model and L: is the log-likelihood. The model with the best fit has the lowest AIC value compared to the others. When comparing models that are not nested and were fitted to the same data set using maximum likelihood, AIC is utilized. (ADETI , 2016)

2.7.2 The Bayesian Information Criterion

One of the most common and often used tools for choosing statistical models is the Bayesian information criterion (BIC). The probability that BIC will choose the true model increases with the size of the training dataset according to the Bayesian probability application's derivation of BIC, which states that if a selection of candidate models includes a true model for the dataset. For the AIC score, the same cannot be true.

BIC=-2logLikelihood +2*logN*k …. (2.28)

Where: N indicates the number of taken observations, and k represent the estimated parameter number.

Calculating the BIC for each model is all that is necessary to compare them using the Bayesian information criteria; the model with smallest BIC is considered to be the best model.(LEE & JOHN, 2003)

3. Results and Discussions:

in this section we explain the basic idea about using both statistical methods to study the most important factors affecting patients with stomach cancer in in Erbil city and their ability to survive and compare between them. The research includes an applied analysis of survival using (Cox regression and Poisson regression) models to identify the factors affecting on the patient in survival analysis. and then the comparison between them to find out the best model for data analysis for patients with stomach cancer, two statistical measures (BIC and AIC) were used to evaluate the best survival model for this data by using three statistical programs analysis (SPSS V. 25, Mat-lab V. and STATGRAPHICS V. 19).

3.1 Data Collection

The data used in this research was obtained from the official database of the main NanaKali Hospital, where these data were collected by patients through direct contact between the specialist doctor and patients. In this study, data were collected for 375 patients with stomach cancer at NANAKALI Main Hospital for Cancer in the Kurdistan Region of Iraq - Erbil. Data were collected during (5) years. Starting from January 1, 2016 until December 31, 2020 for all stomach cancer patients between the ages of (2 - 99) years and of both sexes (males 215 and females 160). During the study period (243) patients died and (132) survived under censored, with a follow-up period until August 10, 2021, the survival time was measured in months and the data contained (12) variables shown below:

Table 3-1 The Response Variables Measured for these Data at Diagnosis:

Variable Name	Description
Age	Age of patient at diagnosis stomach cancer
Gender	Female = (1), Male = (2)
Event status	Alive = (1), Died = (2)
Morphology	Adenocarcinoma = (1), Atypical carcinoid tumor = (2), B- cell lymphoma = (3), Carcinoid tumor = (4), Carcinoma = (5), Gastrointestinal stromal tumor = (6), Hodgkin lymphoma = (7), Lienitis plastic = (8), Lymphoma = (9), Malignant lymphomas = (10), Mucinous adenocarcinoma = (11), Neuroendocrine carcinoma = (12), Signet cell ring adenocarcinoma = (13), Tubular adenocarcinoma = (14)
Behavior	Uncertain = (1), In situ = (2), Malignant = (3)
Grade	grade I = (1), grade II = (2), grade III = (3), grade IV = (4), B-Cell = (5), Un known = (6)
Extent	Localized = (1), Regional by direct extension = (2), regional lymph nodes = (3), regional direct extension and lymph nodes = (4), distant metastasis = (5), not applicable = (6), un known = (7)
Surgery	Does not make surgery = (0), Made surgery = (1)
Radio	Does not take Radiotherapy = (0), Took Radiotherapy= (1)
Chemo	Does not inject Chemotherapy = (0), injected Chemotherapy = (1)
Hormone	Does not use hormone = (0), Used hormone = (1)
Immune	Does not take immune system = (0), Took immune system = (1)

3.2 Application of Cox-Proportional Hazard model

One of the best methods for measuring the patients' ability to live for a specified period of time after medication is the application of Cox regression. In clinical studies, the effectiveness of an intervention is measured by counting how many individuals lived or were saved as a result of that intervention over time. The model building process in this study occurs in eleven variables (Gender, Morphology, Behavior, Grade, Extent, Surgery, Radio, Chemotherapy, Hormone, Immune, Age group)

Table 3-2 Case Processing Summary in Cox-PH Available in Analysis

Case Processing Summary

		N	Percent
Cases available in analysis	Event a	243	64.8%
	Censored	132	35.2%
	Total	375	100.0%
Total		375	100.0%

a. Dependent Variable: time

Table 3-2 shows the case processing summary in Cox PH available in the analysis that determines whether the event occurred for a specific case or not, the number of cases available in the event analysis is 375 cases, the analysis shows that there are 243 deaths, 64.8% are event data and 132 cases, 35.2%, is the number of patients who are still alive under observation because the event did not happen to them.

Omnibus tests are a type of statistical test for all variables, sometimes called the chi-square test. It is a statistical test carried out on a general hypothesis that tends to find general significance between the variance of parameters, while checking parameters of the same type they test whether the variance shown in the set of data is much larger than the unexplained variance in general. The hypothesis is:

H_0: The model includes explanatory variables.

H_1: The model not includes explanatory variables.

Table 3-3 Omnibus Tests of Model Coefficients

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	D.f	P-value	Chi-square	D.f	P-value	Chi-square	D.f	P-value
2493.349	57.095	11	.000	54.651	11	.000	54.650	11	.000

Table 3-3 shows that the value of chi-square = 57.095 at the degree of freedom of 11 and the value of P-value is less at the level of significance 0.05, which means that the statistical model is statistically significant, which indicates that the variables in the model have importance and effect. Thus, we accept the null hypothesis, which states that the explanatory variables are included in the statistical model.

Table 3-4 Omnibus Tests of Model Coefficients by Forward Stepwise (Conditional LR)

step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	Df	P-value	Chi-square	Df	P-value	Chi-square	Df	P-value
1[a]	2533.201	16.495	1	.000	14.798	1	.000	14.798	1	.000
2[b]	2526.978	23.308	2	.000	6.223	1	.013	21.021	2	.000
3[c]	2519.974	30.197	3	.000	7.004	1	.008	28.025	3	.000
4[d]	2514.073	36.593	4	.000	5.902	1	.015	33.926	4	.000
5[e]	2508.404	43.009	5	.000	5.669	1	.017	39.595	5	.000
6[f]	2499.964	50.358	6	.000	8.440	1	.004	48.035	6	.000

Table (3-4) We also note 2-Log Likelihood that the result before including variables within the model chi-Square = 2533.201 and after including variables, the result was 2499.964 for 2-Log The probability of logging this decrease confirms the effect and contribution of the variables to the model. This indicates that the model is statistically significant.

Table 3-5 Variables in the Equation for Cox Regression

Variables in the Equation

x	B	SE	Wald	df	p-value	Exp(B)	99.0% CI for Exp(B)	
							Lower	Upper
Gender	.169	.134	1.601	1	.206	1.185	.839	1.673
Morphology	.058	.014	18.379	1	.000	1.060	1.024	1.098
Behavior	.949	.427	4.934	1	.026	2.582	.860	7.756
Grade	.123	.051	5.861	1	.015	1.131	.992	1.289
Extent	.103	.037	7.919	1	.005	1.109	1.009	1.218
Surgery	-.422	.150	7.892	1	.005	.656	.445	.966
Radiotherapy	.545	.175	9.659	1	.002	1.725	1.098	2.711
Chemotherapy	.341	.209	2.646	1	.104	1.406	.820	2.411
Hormone	.345	.550	.393	1	.531	1.412	.342	5.826
Immunol	-.675	.449	2.258	1	.133	.509	.160	1.619
Age (Binned)	.016	.042	.142	1	.706	1.016	.911	1.133

Table 3-5 shows estimates of the model's coefficients, standard error and degree of freedom, in addition to Wald's test. It also shows the covariates within the model that have no effect or effect by comparing the value of the covariate with the other categories for each of the variables using Exp (B) are called hazard ratios (HR), shows that the event hazard increases as the value of the ith covariate increases, and therefore the duration of survival decreases., if it is equal one, that is, there is no effect on the event,

in summarize:

- HR = 1: No effect
- HR < 1: decrease in the hazard
- HR > 1: Increase in Hazard

and based on the level of significance, the value of 0.05 to achieve the hypothesis results, where the value of the covariate is greater than one. P-value of the variable is greater than the value specified at the level of significance 0.05 indicates the variable has no effect on the event, and if the value of P-value of the variable is less than the value specified at the level of significance 0.05 indicates the effect of the variable on the event and the patient's survival time.

We explain each variable and their effects on patients with stomach cancer as follows:

- Morphology is considered as one of the variables that have an impact on increasing the event risk of the patient's survival, a value of Exp(B) = 1.060, which is an increase in the risk of death for to the patient and p-value = 0.000 which is statistically significant, which indicates a significant effect on the stomach cancer patient. This variable increases the likelihood of death.

- behavior variable is considered a factor in decreasing the survival of the patient with stomach cancer with a value of Exp(B) = (2.582). furthermore, the significant of P-value = 0.026 less than 0.05, this indicates that the variable is statistically significant. In a way that negatively affects the patient's survival, which increases the risk of death of the patient.
- the value Exp(B) for the Grade of cancer is equal 1.131 from the value of B = 0.123, which indicates a significant effect on the patient with stomach cancer, this factor increases the risk of death of the patient. By noting a significant value P-value = 0.015 ≤ 0.05, which is statistically significant.

- The value of Exp(B) for the extent of stomach cancer is equal 1.109 means that, considered to be a factor that has effect on increasing in the patient risk of death in stomach cancer. with p-value = 0.005 which is statistically significant effect on stomach cancer.
- Another factor is surgery. The value of Exp(B) for surgery means that the stomach cancer hazard for all patients that make a surgery are 0.656 months. Surgical operations to remove cancerous tumors is considered one of the factors that have an effect, which decreases the risk of patient's death and the p-value 0.005 < (α = 0.05), which is statistically significant.
- According to our data the radiotherapy has an effect on survival of patients The estimated risk in the radiation group is Exp(B) = 1.725, which is an increase in the risk of death for patient to have or haven't radiotherapy. the p-value is 0.002 and is statistically significant.
- Gender, Chemotherapy, Hormone, Immunol and Age group, are not significant factors because their p-value are greater than (0.05).

When x is the vector of all the fixed variables (Gender, Morphology, Behavior, Grade, Extent, Surgery, Radiotherapy, Chemotherapy, Hormone, Immunol, Age-binned) and β is the corresponding vector of the regression coefficient for the fixed covariates.

The Cox-PH model with significant factor as follows:

$$h_i(t) = h_0(t)\exp(0.058\,Morphplogy + 0.949\,Behavior + 0.123\,Grade + 0.103\,Extent - 0.422\,Surgery + 0.545\,Radiotherapy)$$

3.3 Application of Poisson regression

Poisson regression was applied as a statistical tool in our study because it uses with countable data, any non-fractional integers. Thus, we built a statistical model using Poisson regression to estimate the relationship between a response variable (event) and multiple variables and the extent to which these variables affect the Stomach cancer patient, and knowing the patient's ability to survive.

Our current study recorded eleven explanatory variables (Gender, Morphology, Behavior, Grade, Extent, Surgery, Radio, Chemotherapy, Hormone, Immune, Age group) considered to have an effect on the response variable representing patient survival (event).

A likelihood ratio test is used to determine whether the independent variables in total make the model more accurate than the intercept-only model (i.e., with no independent variables added). With the independent variables present in our example model, we have a p-value of.000 (i.e., p =.000), indicating that the model as a whole is statistically significant, as shown in table (3-6).

\

Table 3-6 Omnibus Tests of Model Coefficients

Likelihood Ratio Chi-Square	df	Sig.
55.250	11	.000

Dependent Variable: status

Table 3-7 summarizes the effect of each predictor, the signs of the coefficients for covariates and the relative values of the coefficients for component levels can provide crucial information about the effects of the predictors in the model. For covariates, positive (negative) coefficients indicate positive (inverse) relationships between the predictors and the outcome. The increased value of the positive covariate corresponds to the rate of reduction in the severity of the event, which is the patient's survival p

we can make the following interpretations based on parameter estimates.

Table 3-7 Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	6.170	.112	5.954	6.386	3133.996	1	.000
gender	-.108	.017	-.140	-.075	41.738	1	.000
Morphology	-.022	.002	-.026	-.019	127.184	1	.000
Behavior	-.474	.030	-.533	-.414	242.144	1	.000
Grade	-.061	.007	-.075	-.047	73.683	1	.000
Extent	-.057	.004	-.066	-.048	156.941	1	.000
Surgery	.073	.019	.035	.111	14.129	1	.000
Radiotherapy	-.205	.025	-.254	-.156	66.924	1	.000
Chemotherapy	-.239	.023	-.285	-.194	105.849	1	.000
Hormone	-.051	.067	-.184	.082	.556	1	.456
Immunol	.221	.045	.132	.310	23.901	1	.000
Age (Binned)	-.026	.005	-.036	-.015	24.407	1	.000
(Scale)	1a						

The Poisson regression model was applied at a significant level of α = 0.05. Table (3-13) shows the parameters, estimates, standard errors, z-values, and p-values from the Poisson regression model. The output shows the results of the fit of the Poisson regression model to describe the relationship between patient survival time and 12 independent variable(s). It turns out that the P-value is less than 0.05, there is a statistically significant relationship between the variables at the 95.0% confidence level. In addition, the P-value of the residuals is greater than or equal to 0.05, indicating that the model is not significantly.

According to the results of the Wald Chi square test, From the table (3-13), it shows that (Gender, Morphology, Extent,

Behavior, Surgery, Chemotherapy, Immunol, Age) estimated coefficient are significant because their (p-values) are less than 5% level, with intercept too, that contribute negatively or positively to survival Patient, while the estimated coefficient of the variable (Hormone) is not significant because it is p-values is greater than the 5% level. we can make the following interpretations based on parameter estimates.

- (Intercept) – This is the Poisson regression estimate when all variables in the model are evaluated at zero, the log of the expected count for the response variable is 6.170 units.
- According to our data, the coefficients of variables (Gender, Morphology, Extent, Behavior, Chemotherapy, Age) are significant in poisson regression model and have an effect on survival of patients. The estimated risk of all of them are greater than one (Exp(B) >1), which is means an increase in the risk of death for patients.
- The results of estimated coefficients of poisson regression showed that the variables (Surgery, Immunol) are considered two factors that have an effect which decreases the risk of patient's death.
- In regards to the variable, Hormone, the p-value is (0.456) greater than (0.05). We would fail to reject the null hypothesis and conclude the Poisson regression coefficient for Hormone is not significant.

Furthermore, we can write the Poisson regression equation with just significant variables:

$\mathrm{Ln}(y)$ = (6.170 – 0. 108 Gender – 0. 022 Morphology – 0. 474 Behavior – 0. 061 Grade – 0. 057 Extent + 0. 073 Surgery – 0. 205 Radiotherapy – 0. 239 Chemotherapy + 0. 221 Immunol – 0. 026 Age-binned)

3.4 Comparing models

To identify the best distribution for the error terms, we will also compare the model summary statistics such as the Akaike information criterion and the Bayesian information criterion and -2 log(likelihood). The best model between Cox regression and Poisson regression will have the lowest values for all 3 statistics. The results are obtained using MATLAB in the following table:

Table 3-8 comparing models with AIC and BIC

Models	No. of parameters	Log Likelihood	AIC	BIC
Cox regression	11	1273.999	1298	1339.2
Poisson regression	11	782.982	806.982	848.1782

Table (3-8) indicates the results for the AIC and BIC values which are used to comparing between two models, which of the two models is more suitable in our data (Cox-model or Poisson Model), the best model should display the lowest AIC, BIC and -2log(likelihood) values.

The results show that Poisson regression model is the best model for our study data of stomach cancer because, it's AIC equals to 806.982and BIC equals to 848.1782 are the lowest values in comparison to the Cox regression model's AIC equals 1298 and BIC equals 1339.2.

2. Conclusion

1. The Omnibus test of model effects for Poisson and cox models have demonstrated that the model fits the chosen variables however when their p-values are smaller than (0.05), which means that the statistical model is statistically significant, which indicates that the variables in the model have importance and effect.
2. by the value of p-value of the Wald Chi square test improved statistically significant variables it shown that for both models Poisson and cox models are not identified the same prognostic factors that influenced in stomach cancer for our data set.
3. According to the results of the cox model, the most significant variables that have an impact on stomach cancer disease are (Morphology, Behavior, Grade, Extent, Surgery, Radiotherapy).
4. The results of Weibull model Shows that the variables that effecting on the stomach cancer for our data set are (Gender, Morphology, Behavior, Grade, Extent, Surgery, Radio, Chemotherapy, Hormone, Immune, Age group).
5. the condition of (over dispersion) in the Poisson model cannot gives adequate results. After fitting Poisson and cox models to the data of stomach cancer although there is over dispersion in our data find out that cox model is more performance and capable for our data than Poisson Regression model by using (Log Likelihood, AIC and BIC).

1- References

1. ADETI , F., 2016. Modelling count out comes from dental caries in adults: A comparision of completing statistics models. *Kwame Nkrumah University of Scince and technologu,Kumas,* pp. 59-60.
2. AGRESTI, A., 2006. *An Introduction to Categorical Data Analysis.* s.l.:2007 John Wiley & Sons, Inc.
3. AHMAD, S. M., 2019. Predicting Cancer Survival Patients using Wavelets with Cox Regression Model. p. 41.
4. ALJAS, E., 1988. A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model.. *Journal of the American Statistical Association, 83(401),* pp. 204-212.
5. AMERICAN CANCER SOCITY, 2022. *cancer A-Z.* [Online]
Available at:

https://www.cancer.org/cancer/stomach-cancer/about/what-is-stomach-cancer.html#references

6. Anon., 2016. Poisson regression. *NCSS, LLC. All Rights Reserved.*, pp. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf.
7. CAMERON, A. C. & TRIVEDI, P. K., 2012. Regression analysis of count data. In: *Event history analysis with R.* s.l.:Cambridge university press.Brostrom.
8. C. D., 2003. Modling Survival Data For Medical Research. London Uk Chapman Hall.
9. CONSUL, P. C. & FAMOYE, F., 1992. Generalized Poisson regression model. pp. 89-109.
10. Coxe, S., West, S. G. & Aiken, L. S., 2009. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, pp. 121-136.
11. FOX, J., 2014. Introduction to Survival analysis. *sociology 761*.
12. HOUT, A. V. D., 2017. *Multi-State Survival Models for Interval-censored data.* London: Taylor & Francis Group CRC Press.
13. I. K., 2019. Overall And Relative Survival For Cancer Patients. Unverstetet I Stavanger, P. 7
14. KLEINBUM , D. G. & KLEIN, M., 2012. Survival analysis. In: *A self-learning text (3rd ed.).* New york: Springer.
15. KOROSTELEVA, O., n.d. Survival analysis. In: s.l.:s.n., p. 60.
16. LEE, E. T. & JOHN, W. W., 2003. Statistical methods for survival data analysis. p. 476.
17. LOKESHMARAN A, &. ,. R. E., 2013. BAYESIAN VARIABLE SELECTION FOR COX'S REGRESSION MODEL. *Asia Pacific Journal of Research*, pp. 11 - 23.
18. Montgomery, D. C., Peck, E. A. E. A. & Vining, G. G., 2006. *Introduction to Linear Regression Analysis (4th ed.). Hoboken: John Wiley & Sons.* s.l.:JOURNAL NAME: Engineering, Vol.6 No.12, November 13, 2014.
19. MOREAU, T., O'QUIGLEY, J. & MESBAH, M., 1985. A Global Goodness-Of-Fit Statistic for the Proportional Hazards Model. *Journal of the Royal Statistical Society.*, pp. 212-218.
20. PARZEN , M. & LIPSITZ, S. R., 1999. A Global Goodness-of-Fit Statistic for Cox Regression ModelS. *Biometrics*, pp. 580-584.
21. SCHMIDT, P. & WITTE, A. D., 1998. *Predicting Recidivism Using Survival Models.* 1st ed. London: Springer verlag.
22. SHINGLETON, J. S., 2012. CRIME TREND PREDICTION USING REGRESSION MODELS FOR SALINAS. *MONTEREY*, pp. 22-23.
23. SINGER, J. D. & WILLETT , J. B., 1991. *Using Survival Analysis When Designing and Analyzing Longitudinal Studies of Duration and the timing of Events.* s.l.:Psychological Bulletin.
24. WEI, L. J., 1984. Testing Goodness of Fit for Proportional Hazards Model with Censored Observations. *Journal of the American Statistical Association*, pp. 649-652.
25. فتيحة، ب.، 2015. تقدير مدة البحث عن الشغل لحاملي شهادات تكوين المهني باستعمال نموذج الأخطار النسبية. *مجلة العلوم الاقتصاد والتسيير والتجارة (جامعة بغداد)*، pp. 11-33.