Biomedical Statistics and Informatics
2020; 5(1): 14-19
http://www.sciencepublishinggroup.com/j/bsi
doi: 10.11648/j.bsi.20200501.13
ISSN: 2578-871X (Print); ISSN: 2578-8728 (Online)

Estimation of Population Based Colorectal Cancer Survival Analysis Using Cox Proportional Hazards Model

Marafa Haliru Muhammad[1, *], Usman Umar[1, 2]

[1]Planning Division, Usmanu Danfodiyo University Teaching Hospital, Sokoto, Nigeria

[2]Department of Mathematics, Usmanu Danfodiyo University, Sokoto, Nigeria

Email address:
halirumarafa5@gmail.com (M. H. Muhammad), Uusman07@gmail.com (U. Umar)

[*]Corresponding author

To cite this article:
Marafa Haliru Muhammad, Usman Umar. Estimation of Population Based Colorectal Cancer Survival Analysis Using Cox Proportional Hazards Model. *Biomedical Statistics and Informatics*. Vol. 5, No. 1, 2020, pp. 14-19. doi: 10.11648/j.bsi.20200501.13

Received: December 20, 2019; **Accepted**: December 30, 2019; **Published**: February 4, 2020

Abstract: Colorectal cancer (CRC) is a tumour of the colon and rectum. Most cases of CRC are sporadic; meaning there are no known hereditary (genetic) components, and it develops slowly over several years through adenomatous polyps. Changes in bowel habits, blood in the stool, and anaemia are cardinal symptoms and sings of CRC. In later stages, fatigue, anorexia, weight loss, pain, jaundice, and other signs and symptoms of locally advanced and metastatic disease occur. The aim of this study is to estimate the population based colorectal cancer survival analysis using cox Proportional Hazards model, in order to fits colorectal cancer data in population-based research. This research was a five-year retrospective study on data from a record of colorectal cancer patients that received treatments from 2013 to 2017 in Radiotherapy Department of Usmanu Danfodiyo University Teaching Hospital, Sokoto, being it one of the cancer registries in Nigeria. 9 covariates were selected to fit colorectal cancer data using Cox Regression Models. The 5-year median survival was found to be 121 days. From the results, it was concluded that the predictor variables could significantly predict the survival of colorectal cancer patients using Cox proportional model. Also the results show that the data met Cox Proportional Hazards Assumptions.

Keywords: Colorectal, Cancer, Cox, Hazards, Assumptions

1. Introduction

Colorectal cancer (CRC) is a tumour of the colon and rectum. Most cases of CRC are sporadic; meaning there are no known hereditary (genetic) components, and it develops slowly over several years through adenomatous polyps (Brenner *et al*., [1]). Changes in bowel habits, blood in the stool, and anaemia are cardinal symptoms and sings of CRC. In later stages, fatigue, anorexia, weight loss, pain, jaundice, and other signs and symptoms of locally advanced and metastatic disease occur. CRC is traditionally diagnosed by sigmoidoscopy and colonoscopy using biopsy. There are several ways to treat colorectal cancer depending on the cancer stage and where the tumour is localized. The main treatment is surgery; however, chemotherapy and radiation therapy can also use (Potter & Hunter, [2]).

Approximately 1.4 million new cases of colorectal cancer and almost 700 000 deaths occurred worldwide in 2012 (Arnold *et al.,* [3]). Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc. For example, if the event of interest is death, then the survival time can be the time in years until a person dies (Hosmer D. W., Lemeshow S., and May S., [4]).

According to Hosmer *et al.* [4] observations are called censored when the information about their survival time is incomplete; the most commonly encountered form is right censoring. Censoring is an important issue in survival analysis, representing a particular type of missing data. Censoring that is random and non-informative is usually required in order to avoid bias in a survival analysis.

The survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. The survival function gives, for every time, the

probability of surviving (or not experiencing the event) up to that time. The hazard function gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time. While these are often of direct interest, many other quantities of interest (e.g., median survival) may subsequently be estimated from knowing either the hazard or survival function (Hosmer *et al.,* [4]).

Many countries today have population-based cancer registries. Their task is to collect and store information on all cases of cancer in the countries and produce statistics of the incidence of cancer, and the survival of cancer patients. They play an important role in analysing the impact of cancer in the community. In Nigeria, for example, there are ten (10) population-based cancer registries owned by the Federal Government located at various tertiary hospitals across the country, according to Nigerian National System of Cancer Registries (NSCR, [5]). In most part of Africa, cancer burden is under reported due to lack of or inaccurate population statistics, which makes age specific incidence rate impossible or inaccurate (Abdulkareem, [6]).

This study was to estimate the population based colorectal cancer survival analysis using Cox proportional hazard model, in order to fits colorectal cancer data in population-based research.

The leading cause of death and disabilities worldwide is cancer which affects more than 14 million people annually (W. H. O., [15]). Knut *et al*. [16] consider colorectal cancer (CRC) as a complex disease that almost 40% of the surgically cured patients experience cancer recurrence within 5 years. Cancer control refers to all actions taken to reduce the frequency and impact of cancer (Armstrong, [17]).

Zaki [18] found a general formula for generating survival data on the computer trough the fundamental relation between hazard rate and survival function. The development of methods in analyzing survival data is one of the areas in statistics that have increased recently.

Nigeria contributed 15% to the estimated 681,000 new cases of cancer that occurred in Africa in 2008 (Sylla, [19]). Similar to the situation in the rest of the developing world, a significant proportion of the increase in incidence of cancer in Nigeria is due to increasing life expectancy, reduced risk of death from infectious diseases, increasing prevalence of smoking, physical inactivity, obesity as well as changing dietary and lifestyle patterns (Sylla, [19]).

2. Material and Method

This research was a five-year retrospective study on data from a record of colorectal cancer patients that received treatments from 2013 to 2017 in Radiotherapy Department of Usmanu Danfodiyo University Teaching Hospital, Sokoto. A purposive sampling was considered in selecting UDUTH being it one of the cancer registries in Nigeria.

The research was designed to follow the subsequent procedure. The first stage was the discussion and formulation of Cox Proportional-Hazards Model. Finally, the data from one of the cancer registries (Usmanu Danfodiyo University Teaching Hospital, Sokoto) were collected for the following estimates: Kaplan-Meier Plots, test survival curves using Log-rank tests (Survival, Hazard and Median Survival Functions).

Software: The *R* programming language has sufficient packages required to carry out the research work. And SPSS was used for data entries and arrangements.

2.1. Kaplan Meier

In cancer trial, Kaplan-Meier (K-M) method is one of the recommended techniques in survival analysis: it is the most popular in developing survival function (Collett, [20]). The method is used to measure the fraction of subjects living for a certain period of time after treatment. It is applied by analyzing the distribution of patients' survival times following their recruitment to a study. The analysis expresses in terms of proportion of patients still alive up to a given time, following their recruitment. In terms of graph, a plot of proportion of patients' surviving against time has a characteristic decline; the steepness of the curve indicates the efficacy of the treatments being investigated. The shallower part of the curve shows the more effective treatment. In analysing the survival data, two functions that are dependent on time are of particular interest: the survival function and the hazard function.

The survival function denoted by S (t) is the probability of surviving at least to time *t*.

The hazard function denoted by h (t) is the conditional probability of dying at time *t* having survived to that time. The graph of S (t) against *t* is called the survival curve.

The Kaplan-Meier method can be used to estimate this curve from the observed survival times without the assumption of the underlying probability distribution. The method is based on the basic idea that the probability of surviving p or more periods from entering the study is the product of the p observed survival rates for each period i.e. the cumulative surviving, and is given by

$$S(p) = (k_1)(k_2) \dots (k_p) \quad (1)$$

where

k_i Denotes the proportion of surviving the ith period

$i = 1,2, \dots p$ = Proportion of surviving beyond the second period conditional on having survived up to the second period and so on.

The proportional surviving period i having survived up to period i is given by

$$K = \frac{r_i - d_i}{d_i} \quad (2)$$

where

r_i = the numbers alive at the beginning of the ith period

d_i = The number of deaths within the ith period

2.2. Log-rank Test

A statistical hypothesis test called Log-rank test was used to compare the two survival curves. It is used to test the null

hypothesis that there is no difference between the survival curves, i.e. the probability of event occurring at any point in time is the same for each population.

The total expected number of events for a group was the sum of the expected number of events at the time of each event. The expected number of events at the time of an event can be calculated as the risk for death at that time multiplied by the numbers alive in the group. Under the null hypothesis, the risk of death, i.e. number of deaths divided by the numbers alive can be calculated from the combined data for these groups.

$$E_2 = \sum_{i=1}^{k} \frac{d_i}{r_i} r_{2i} \quad (3)$$

where

r_{2i} = the numbers alive from group 2 at the time of event i

E_1 is calculated as $n - E_2$, where n = the total number of events

The test statistic is compared with a $\chi^2 - distribution$ with 1 degree of freedom.

2.3. Cox Proportional-hazards Regression

The Proportional Hazards Model, proposed by Cox [21], has been used primarily in medical testing analysis to model the effect of secondary variables on survival. Its strength lies in its ability to model and test many inferences about survival without making any specific assumptions about the form of the life distribution model.

Most interesting survival-analysis research examines the relationship between survival typically in the form of the hazard function and one or more explanatory variables (or *covariates*).

The most common are linear-like models for the log hazard. For example, a parametric regression model based on the exponential distribution:

$$log_e h_i(t) = h_o(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (4)$$

Or equivalently,

$$h_i(t) = h_o(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \quad (5)$$

$$= h_o(t) \times e^{\beta_1 x_{i1}} + e^{\beta_2 x_{i2}} + \cdots + e^{\beta_k x_{ik}} \quad (6)$$

where

$h_i(t)$ = Denotes the Hazards Function

$h_o(t)$ Is the Baseline Hazards

β_i Represents the Relative Risk

x_{ij} Represents the Covariates

$$i = 1, 2 \ldots N$$

$$j = 1, 2 \ldots k$$

Where i are indexes subjects, $x_{i1}, x_{i2} \ldots x_{ik}$ are the values of the covariates for the i^{th} subject.

This is therefore a linear model for the log-hazard or a multiplicative model for the hazards itself. The model is *parametric* because, once the regression parameters $h_o(t), \beta_{1,\ldots,}\beta_k$ are specified, the hazard function $h_i(t)$ is fully characterized by the model, the regression constant represents a kind of *baseline hazard* when all of the $x's$ are 0. Other parametric hazard regression models are based on other distributions commonly used in modelling survival data such as the Weibull distributions.

Fully parametric hazard regression models have largely been superseded by the Cox model [21], which leaves the baseline hazard function $h_o(t) = log_e h_i(t)$ unspecified:

$$log_e h_i(t) = h_o(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

or equivalently,

$$h_i(t) = h_o(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \quad (7)$$

The Cox Model is termed *semi-parametric* because, while the baseline hazard can take any form, the covariates enter the model through the *linear predictor*

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (8)$$

Notice that there is no constant term (intercept) in the linear predictor: The constant is absorbed in the baseline hazard. The Cox Regression Model is a *Proportional-Hazards Model*:

Consider two observations, i and i', that differ in their x-values with respective linear predictors

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (9)$$

And

$$\eta_{i'} = \beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_k x_{i'k} \quad (10)$$

The hazard ratio for these two observations is

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_o(t) e^{n_i}}{h_o(t) e^{n_{i'}}} = \frac{e^{n_i}}{e^{n_{i'}}} = e^{n_i - n_{i'}} \quad (11)$$

This ratio is constant over time. In this initial formulation, the research assumed that the values of the covariate x_{ij} are constant over time.

As we will see later, the Cox model can easily accommodate *time-dependent covariates* as well.

3. Results of Findings

3.1. Results of Kaplan Meier Plot without Covariates

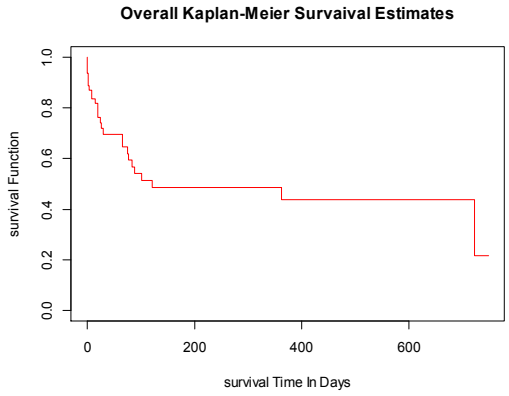

Figure 1. *Kaplan-Meier (K-M) Curve for overall survival estimate.*

Figure 1 Show that the overall median survival time is 121 days. This implies that 50% of the colorectal cancer patients survived less or equal to 121days and the other 50% survive longer than 121days after they are diagnosed with the disease. This is the survival time at which the cumulative survival function is equal to 0.5.

3.2. Results of Kaplan Meier Estimates with Covariates

Table 1. *Results from K-M Plots.*

Covariate	Categories	Median Survival
Age	1-20	121
	21-40	
	41-60	102
	61-80	88
Associated Comps.	No Comps	361
	HBP	2
	DM	20
Family History	Yes	88
	No	
Sex	Male	121
	Female	77
Stage	Non-Specific	66
	Stage A	
	Stage B	
	Stage C	102
	Stage D	
Tribe	Hausa	121
	Yoruba	
	Igbo	
	Ibra	
	Igala	
	Nupe	15
	Non-Specific	76
Type of Colorectal	Colonic	
	Rectal	361
	Sigmoid	121
Type of Treatment	Single	121
	Combine	

3.3. Results from Log-rank Test

Table 2. *Results from Log-Rank Tests.*

Covariates	D. F	Log-Rank Test	P
Age	3	0.6	0.9
Age At Diagnosis	3	0.6	0.9
Associated Complecations	2	14.4	0.0007
Family History	1	4.5	0.03
Sex	1	0.1	0.8
Stage	4	11.9	0.02
Tribe	5	4.7	0.5
Type Of Colorectal	3	5.9	0.1
Type Of Treatment	1	0.6	0.4

3.4. Results from Cox Proportional Hazard Model

Table 3. *Results from Cox Proportional Hazards Model.*

Covariates	Coef	Exp (coef)	Se (coef)	Z	P	S
Age	-19.290	0.000	0.009	-2038.04	0.000	***
Age At Diagnosis	19.290	238600000	0.009	2038.049	0.000	***
Associated Complecations	0.122	1.130	0.299	0.410	0.682	
FamilyHistory	-20.000	0.000	10540	-0.002	0.998	
Sex	1.139	3.123	0.449	2.535	0.011	*
Stage	-0.871	0.418	0.328	-2.661	0.008	**
Tribe	0.069	1.071	0.276	0.250	0.802	
Type Of Colorectal	-0.405	0.667	0.237	-1.712	0.087	.
Type Of Treatment	0.353	1.423	0.414	0.852	0.394	

Log Likelihood = - 68.097.

From table 3, the Cox Proportional Hazard result shows that the colorectal cancer patients receiving combined therapy (surgery and chemotherapy) have higher risk of death event than those receiving single therapy, and increasing the covariate Type of Treatment with 1 unit will increase the hazard ratio by 0.353. So, it is not significant since the p-value = 0.394. The result of the covariate Sex shows that it is significant, and increasing the Sex by 1 unit will increase the hazard ratio by 1.130. This indicates that the female patients have the higher risk of death than the male patients. Increasing the covariate Age by 1 unit will decrease the hazard ratio by -19.290; so it is highly significant.

3.5. Results from Cox Proportional Hazards Assumption Using Statistical Test

Table 4. *Results from Cox Proportional Assumptions Using Statistical Test.*

Covariates	rho	Chisq	P
Age	-0.08891	0.1690	0.6811
Age At Diagnosis	-0.08891	0.1690	0.6811
Associated Complecations	-0.00137	0.0000	0.9946
Family History	-0.50605	0.0000	0.9999
Sex	0.14503	0.5870	0.4435
Stage	0.19202	0.9660	0.3256
Tribe	-0.04515	0.0984	0.7538
Type Of Colorectal	-0.30827	4.1700	0.0511
Type Of Treatment	0.01961	0.0095	0.9222
GLOBAL	NA	6.1700	0.7226

From table 4 the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards (which mean that proportion hazards assumptions are met).

3.6. Results from Cox Proportional Hazard Assumptions Graphical Method

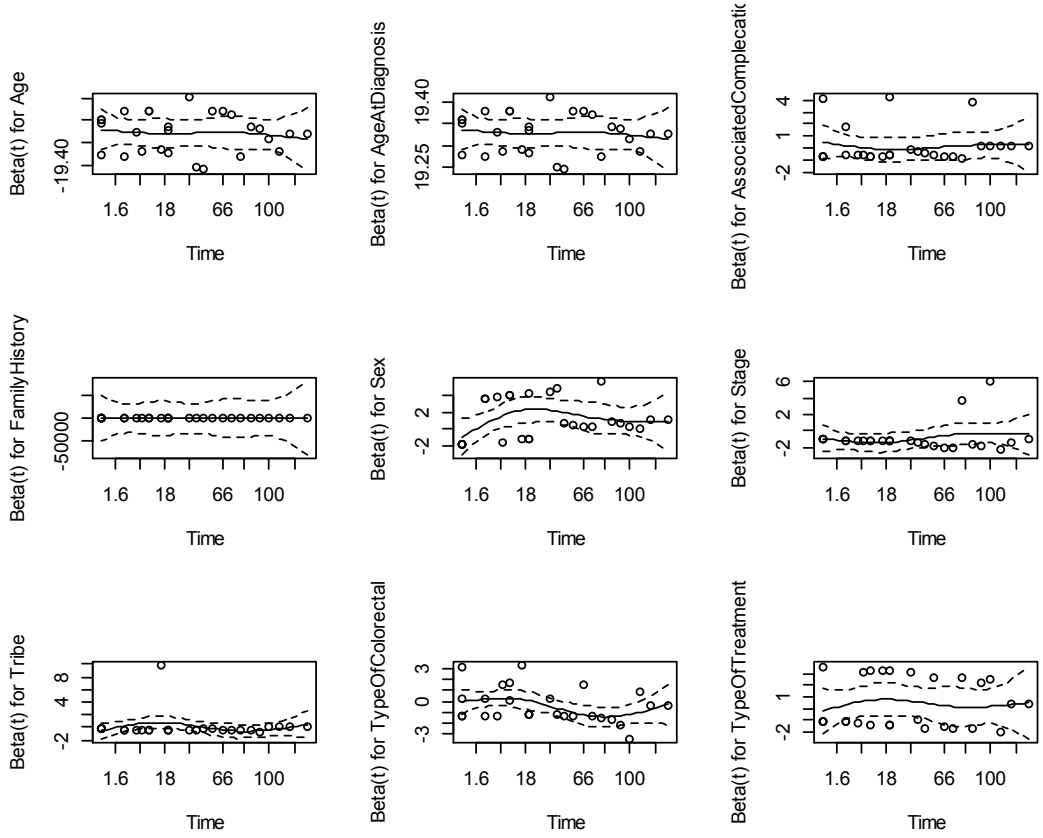

Figure 2. *Cox Proportional Hazard Assumptions Graphical Method.*

In the figures 2, the solid lines of the graphs are the smoothing spline fit to the plot, with the dashed lines representing a standard-error band around the fit. From the graphical inspection, there is no pattern with time. The assumption of proportional hazards appears to be supported for the covariates.

4. Conclusions

The results of this study shows that, according to our colorectal cancer data, the semi-parametric Cox regression model could better determine the factors associated with the colorectal cancer disease. However, in the present study, the Cox model provided an efficient and a better fit to the study data. Therefore, it would be better for researchers of the health care field to consider this model in their researches concerning the colorectal cancer disease if the assumptions of proportional hazards are fulfilled.

References

[1] Brenner, H., Kloor, M., & Pox, C. P. (2014). Colorectal cancer. *The Lancet*, 1490-1502.

[2] Potter, J. D., & Hunter, D. (2008). Colorectal Cancer. In H.-O. Adami, D. Hunter, & D. Trichopoulos, *Textbook of Cancer Epidemiology* (pp. 275-297). New York: Oxford University Press, Inc.

[3] Arnold, M., Sierra, M. S., Laversanne, M., & Soerjomataram, I. (2017). Global patterns and trends in colorectal cancer incidence and mortality. *Gut, pp. 683-691.*

[4] Hosmer D. W., Lemeshow S., and May S. (2008). Applied Survival Analysis: Regression Modeling of Time- to- Event Data. 234-654.

[5] Nigerian National System of Cancer Registries (NSCR) 2018.

[6] Abdulkareem, F., (2009). Epidemiology and Incidence of Common Cancer in Nigeria. Presentation of Cancer Registration and Epidemiology Workshop. April, 2009.

[7] Dickman, P. W. and Hakulinen, T. (2008). Population-Based Cancer Survival Analysis. John Wiley and Sons, UK 23-65.

[8] Kleinbaum D G and Klein M. (2005) *Survival Analysis* New York: Springer. 16-53.

[9] Dickman, P. W. (2010). An Introduction and Some Recent Development in Statistical Methods for Population-Based Cancer Survival Analysis. Statistical Methods for Population-Based Cancer Survival Analysis. Milan, 33-55.

[10] Adejumo, A. O. and Ahmadu, A. O. (2016) A Study of the Slope of Cox Proportional Hazard and Weibull Models. *Science World Journal* Vol 11 (No 3) 2016.

[11] Quantin, C., Michal, A., Thierry, M., Gillian, B., Todd, M., Mohammed, A., et al., (1999) Variation over Time of the Effects of Prognostic Factors in a Population based Study of Colon Cancer: Comparison of Statistical Models. *American Journal of Epidemiology, Vol. 150,* 11-20.

[12] Ahmed E. F., Paul W. V., Don H., (2007) Modeling survival in colon cancer: a methodological review. *Molecular Cancer* 2007, 6: 15 doi: 10.1186/1476-4598-6-15.

[13] Abdulkabir M., Ahmadu A. O., Udokang A. E., Raji S. T., (2015). *Gradient Curve of Cox Proportional Harzard and Weibull Models*. Automatic Control of Physiological State and Function, 2: 2 http://dx.doi.org/10.4172/2090-5092.1000108.

[14] Wang Kesheng, Xuefeng Liu, Yue Pan, Daniel Owusu1, Chun Xu. (2017) Comparison of Cox Regression and Parametric Models. *Journal of Data Science* 16 (2017), 423-442.

[15] World Health Organisazation. 2013a Cancer Control: A Global Snaptshot in 2015http://www.who.int/cancer/cancer-snapshot-2015/en/ Accesed on 9 November 2016.

[16] Knut A. M., Dejan Ignjatovic, Marianne A. M. (2017) Tailored Treatment of Colorectal Cancer: Surgical, Molecular, and Genetic Considerations. *Clinical Medicine Insights: Oncology*. DOI: 10.1177/1179554917690766.

[17] Armstrong, B. K. (1992). The Role of Cancer Registry in Cancer Control. Cancer Causes and Control. 3: 569 – 579.

[18] Zaki A. (2015) *Log-linearity for Cox'sregression model.* University of Oslo.

[19] Sylla BS, Wild CP (2011). A million Africans a Year Dying from Cancer by 2030: What can cancer research and control offer to the continent? Int J Cancer.

[20] Collett, D. (2003). Modelling Survival Data in Medical Research. Chapman and Hall, London, 37-87.

[21] Cox D: Regression Models and Life Tables (with Discussion). *J Roy Stat Soc B* 1972, 4: 187-220.