Asian Pacific Journal of Cancer Care
Vol 7 No 4 (2022), 601-605
Original Research

Comparison of Survival Estimation Methods in the Analysis of Breast Cancer Data

Olubimpe M. Oladuti Department of Statistics, Federal University of Technology, PMB 704, Akure, Nigeria.
Paul O. Olopha Department of Statistics, Federal University of Technology, PMB 704, Akure, Nigeria.

Objective: In many epidemiological studies or clinical trials, Cox Proportional Hazard model is the most commonly used technique for the analysis of effects of covariates on survival time when time are continuous and ties are present. It is important to handle the presence of ties in the data to prevent the inherent effect which may lead to unreliable result in the analysis. Therefore, this study was carried out handle the inadvertent inclusion of ties in survival time with three different estimation methods; Breslow, Efron and exact partial likelihood estimation methods of Cox regression model.
Material and Methods: Analysis were carried out using data collected on 300 breast cancer patients from University of Ilorin Teaching Hospital to evaluate the factors affecting the survival of patients with breast cancer. The estimation methods used in handling ties in the breast cancer data are compared using Akaike Information Criterion. All analyses were performed using STATA version 14.0.
Result: Of a total of 300 patients, 97 (32.33%) died of breast cancer. The result of the data analysis using Cox model with different estimation methods were approximately alike. The risk of dying from the breast cancer was associated with age of the patients; patients with Cytological diagnosis had an increased risk for death.
Conclusion: The hazard ratios of the estimators are approximately similar but based on Akaike Information Criteria, exact partial likelihood outperformed the other methods of handling ties in breast cancer data since it has the smallest value.

Introduction

According to American Cancer Society (2018) [1], Breast cancer is a group of diseases that cause abnormal cells to split without control and overpass other tissues both in men and mainly in women. They came out with a finding that breast cancer is 100 times more common in women than in men and estimates that each year, about 1990 new cases of breast cancer in men will be diagnosed and that breast cancer will cause approximately 480 deaths in men. According to GLOBOCAN 2012, breast cancer is the most common type of cancer for both sexes especially women worldwide with an estimated 14.1 million new cases diagnosed in 2012. Globally, in 2012 among the three most leading cancer; breast cancer is 11.9%. According to World Health Organization [2], 560 thousand women died as a result of breast cancer out of 1.6 million cases diagnosed. The occurrence and prevalence of breast cancer are escalating globally. Although a greater proportion of women are diagnosed in early disease stages because of national screening programs and awareness [3]. According to Azubuike SO et al, 2018 [4], Africa has extremely high mortality due to Breast cancer. Consequently, some expert including the World Health Organization endorse prompt diagnosis couple with apt and effective treatment as a valuable measure for reducing mortality via breast cancer [5].

It was also reported that the incidence and mortality rates of breast cancers are decreasing in developed countries while on the contrary in developing countries [6] which is in agreement with the prediction of WHO (World Health Organization, 2005), that there will be a major increase in cancer incidence and mortality in developing countries as a result of increase in lifespan,

development, and advanced contact to risk factors [2].

Studies have been shown that breast cancer is a challenge globally even with the advances in investigation, diagnosis and treatment. Also, numerous prognostic factors such as old age, hereditary, delay in pregnancy, using oral contraceptive, early menstruation, late menopause and genetic are associated to breast cancer worldwide [2].

The aims of this study was (i) to use Cox regression model to evaluate the factors that affect the survival of patients with breast cancer, (ii) to use exact method, Efron method and Breslow method of approximation to handle the occurrence of ties in the data and (iii) compare the performances of these methods.

In order to determine the best model, Akaike Information Criterion (AIC) was calculated and compared. The rest of the paper is organised as follows; section 2 of the paper describes the materials and methods used in analyzing the data. In section 3, analyses of data using Cox regression model with Breslow, Efron and Exact approximation methods are carried out and the result presented. Finally, discussion and conclusion of the results are presented in section 4.

Materials and Methods

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 survival time is a length of time t that corresponds to the time period between a well-defined start-time t_0and the time t_c of an event. This event can be either death, occurrence of a disease, marriage, divorce or any designated experience of interest that may happen to an individual, and the time to event or survival time can be measured in days, weeks and years. The right censoring is the most common censoring in survival time data; it assumed time to event T* and the right censoring time C for some individuals in the study. The exact survival time T of any individual will be known if and only if $T^* \leq C$. Conversely, if $T^* > C$, the individual is a survivor and the exact survival time is censored at C. Hence, the observed time is $T=\min(T^*,C)$ represented by a pair of random variable (T,δ), where δ signifies whether the survival time T corresponds to an event ($\delta=1$) or right censored ($\delta=0$). Survival time with censoring frequently occurred in many medical and reliability studies [7]. Analyzing censoring data has been a major challenge in medical research. Semi-parametric and parametric models are used for right censored data but Cox proportional hazards model is one of the most commonly used semi-parametric model which does not require any specific assumption about the shape of survival function. It is the most flexible continuous- time model which estimates the relationship between the hazard rate and explanatory variable under the basic assumption that survival time are untied [8] but in actual fact, there is always some smallest time unit that ties can occur if time are measure in a flawlessly continuous scale. Conversely, in many data sets, ties are present typically due to the fact that failure times are continuous and only reported to the nearest day. Cox (1972) [9] proposed a proportional hazards model for event times when the event times are continuously distributed and the possibility of ties is ignored but in the analysis of survival time data, ties between event times are important to consider when fitting the Cox model (Cox, 1972) [9] seeing that the Cox partial likelihood depends largely on the order of events. Often times, survival data contain tied observations; that is, two or more individuals in the dataset share the same time and these need to be taken care of. Although, many research approaches assume that for continuous time; equal survival times of a sample have probability zero but this assumption is violated if many ties are present. The three broadly used methods to treat ties between event times in the Cox proportional hazards model are: the exact partial likelihood method [10], Breslow approximation [11] and Efron approximation [12]. According to Huang and Liu (2007), the ideal method of handling ties under Cox regression model formulation is the exact partial likelihood method [7].

Let Zi (i=1,2,…,m) be the value of covariates for the i^{th} individual, the proportional hazards model for ith individual at time t can be written as:

where $\lambda_o(t)$ (t) is the baseline hazard at time t.

exp ($z_i\beta$) is the effects of covariates on the hazard for the event.

According to Lawless (2003), the Cox partial likelihood function is used to evaluate the estimates of the coefficients, where the partial likelihood is given by

where k is the number of different observed event times, z_j is the covariates vector at time t_i and r_j is the risk set which includes individuals whose observed event time or censoring time is greater than or equal to t_j [13].

Breslow (1975) [11] suggests the summing up of covariate related components for all subjects experiencing the event at a given time t_j and raising the result to a power equal to the number of events tied at t_j . The partial likelihood function that uses this approach is defined as

Nevertheless, if the number of tied events for any time t_j is quite large, this method might not give a good approximation of the partial likelihood function [14].

Efron [12] partial likelihood function is then proposed as an alternative estimator approximated as follows;

The exact method is the discrete of partial likelihood function. It based on the continuous likelihood under the assumption that if there are tied events, that is due to the inaccurate nature of our measurement, and that there must be some true ordering. Then, all possible orderings of the tied events are calculated, and the probabilities of each event are summed [15].

Moreover, if the number of tied events is very small, all three methods give very similar results but if there are no ties, all methods lead to exactly the same results. In order to determine the best estimation, Akaike Information Criteria (AIC) is used. The AIC is a measured of the goodness of fit of an estimated statistical model, estimating the quality of each model relative to each of the other models [16].

AIC = -2log (L) + 2p

where p is the number of model parameter L is the maximum value of the likelihood function for the estimated model *Analysis of Data* The present study used data from 300 breast cancer patients; 275 females and 25 males who were admitted at University of Ilorin Teaching Hospital (UILTH) which

covers a period of five (5) years (2011 to 2016). The record of each patient contained information of variables; length of stay in hospital (in days), sex, age of patients, location of cancer, mode of diagnosis and outcomes which indicate whether the patients is dead or alive. Survival time is defined as length of admission before death occurs, while those who were still alive at the time of data collection were right-censored. The analysis was carried out using Cox regression model with three estimations methods of handling ties.

All analyses were carried out using STATA (version 14.0) statistical software.

Results

A total number of 300 patients (8.33% males) with breast cancer were included in the analysis. Based on descriptive analysis, of the 300 patients, 97 were dead and the others were right censored. 140 (46.67%) of the patients have the cancer on their left breast, 135 (45%) of them on their right breast while the remaining 25 (8.33%) have it on both breasts. Patients whose ages falls between 36-50 years are the most affected age-group followed by 51-65 years. Using Cox model, we compared Breslow, Efron and exact partial likelihood using AIC to know the best estimation method for breast cancer data.

Discussion

According to the result, using the three methods of handling ties for Cox regression model which are presented in Tables 1, 2 and 3, age of the patients has significant effect on mortality rate of patients with breast cancer. Patients in the age groups; <35years, 35-50 and 51-65years are less likely to die from breast cancer compared to those in age group >65years. Patients with cytological diagnosis are more at risk of death due to breast cancer. Although, the hazard ratio in the three estimation methods are approximately similar but from the AIC, Exact partial likelihood (776.666) is better than Breslow (904.165) and Efron (901.621).

Factors	Categories	Hazard Ratio	95% Credible Interval	P-value
Gender	Male **	1		
	Female	0.652	(0.333-1.277)	0.213
Age- Group	<35	0.52	(0.247-1.094)	0.085
	35-50	0.396	(0.223-0.701)	0.001
	51-65	0.523	(0.293-0.931)	0.028
	>65**	1		
Location of Cancer	Left Breast	0.909	(0.464-1.778)	0.78
	Right Breast	0.57	(0.280-1.159)	0.121
	Both Breasts**	1		
Mode of Diagnosis	Histological**	1		
	Cytological	1.046	(0.685-1.597)	0.836
		AIC#	904.1649	

Table 1. Prognostic Factors of Breast Cancer using Cox Regression Model with Breslow Estimation Method.

** Reference category; #Akaike Information Criteria; P-value<0.05

Factors	Categories	Hazard Ratio	95% Credible Interval	P-value
Gender	Male **	1		
	Female	0.647	(0.331-1.268)	0.205
Age- Group	<35	0.513	(0.244-1.079)	0.078
	35-50	0.389	(0.219-0.689)	0.001

	51-65	0.516	(0.290-0.920)	0.025
	>65**	1		
Location of Cancer	Left Breast	0.905	(0.462-1.771)	0.77
	Right Breast	0.562	(0.276-1.145)	0.113
	Both Breasts**	1		
Mode of Diagnosis	Histological**	1		
	Cytological	1.039	(0.681-1.587)	0.858
		AIC#	901.6207	

Table 2. Prognostic Factors of Breast Cancer using Cox Regression Model with Efron Estimation Method.

** Reference category; #Akaike Information Criteria; P-value<0.05

Factors	Categories	Hazard Ratio	95% Credible Interval	P-value
Gender	Male **	1		
	Female	0.646	(0.326-1.280)	0.21
Age- Group	<35	0.51	(0.240-1.082)	0.079
	35-50	0.385	(0.215-0.690)	0.001
	51-65	0.512	(0285-0.922)	0.026
	>65**	1		
Location of Cancer	Left Breast	0.906	(0.458-1.789)	0.776
	Right Breast	0.562	(0.274-1.155)	0.117
	Both Breasts**	1		
Mode of Diagnosis	Histological**	1		
	Cytological	1.047	(0.682-1.606)	0.835
		AIC#	776.6655	

Table 3. Prognostic Factors of Breast Cancer using Cox Regression Model with Exact Partial Likelihood Estimation Method.

** Reference category; #Akaike Information Criteria; P-value<0.05

In conclusion, Cox regression model is a semi- parametric and most frequently used model for the analysis of prognosis factors in clinical research. Perhaps it estimates the relationship between the hazard rate and explanatory variable under the basic assumption that survival times are untied [8] but in actual fact, there is always some smallest time unit that ties can occur. In the analysis of survival time data, ties between event times are imperative to consider when fitting the Cox model [9].

This present study aimed at determining the best estimation method of handling ties in breast cancer data as the right censoring. Three different methods for Cox regression model were considered in this study namely Breslow method, Efron method and exact partial likelihood method. Akaike Information Criterion (AIC) was used to evaluate the performance of each model. Four independent variables used are the prognostic factors on survival of patients for each estimation methods with different estimates. The results also showed significant differences among age groups with respect to the risk of dying.

Ethical Approval

Permission to use data from University of Ilorin Teaching Hospital (UILTH). Respondent

confidentiality was intact as no names and addresses were included in the data set and therefore the respondents cannot be traced by the researcher.

Conflict of Interest

The authors declare no conflict of interest.

References

References

1. American Cancer Society (2018). Cancer Facts & Figures 2014. *American Cancer Society (ACS) Atlanta, GA: American Cancer Society.*
2. WHO. 2015b. Breast cancer: prevention and control [Online]. *Available: http://www.who.int/cancer/detection/ breastcancer/en/index1.html# 4/1/2015].*
3. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. *CA: a cancer journal for clinicians.* 2013; 63(1)DOI
4. Azubuike SO, Muirhead C, Hayes L, McNally R. Rising global burden of breast cancer: the case of sub-Saharan Africa (with emphasis on Nigeria) and implications for regional development: a review. *World Journal of Surgical Oncology.* 2018; 16(1)DOI
5. Silva I, Mccormack V, Jedy-Agba E, Adebamowo C. DOWNSTAGING BREAST CANCER IN SUB-SAHARAN AFRICA: A REALISTIC TARGET?. 2017.
6. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians.* 2021; 71(3)DOI
7. Huang X, Liu L. A joint frailty model for survival and gap times between recurrent events. *Biometrics.* 2007; 63(2)DOI
8. Anderson G, Fleming T. Model misspecification in proportional hazards regression. *Biometrika.* 1995; 82(3)DOI
9. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological).* 1972; 34(2)DOI
10. Kalbfleisch, J. D, Prentice, R. L. The Statisical Analysis of Failure Time Data (2nd ed.). *John Wiley Interscience.* 2002.
11. Breslow N.. Covariance analysis of censored survival data. *Biometrics.* 1974; 30(1)
12. Efron B. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association.* 1977; 72(359)DOI
13. Lawless, J.F. Statistical Models and Methods for Lifetime Data. *New York: John Wiley & Sons Inc.* 2003.
14. Borucka, J. Methods of Handling Tied Events in the Cox Proportional Hazard Model. 2014.
15. Allison P. Survival Analysis Using SAS: A Practical Guide. Cary, NC, USA: SAS Institute Inc. 2010.
16. Akaike H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on.* 1974; 19:716-723.