Research/Review Article

FEDPOLEL JOURNAL OF APPLIED SCIENCES
https://fedpoleljournalofappliedsciences.com/

Comparison of Exact, Efron and Breslow Partial Likelihood Methods for Cox Proportional Hazards Model Parameter Estimation

Adedoyin, E.D.[1] and Awoleye, G.O.[2]

[1]*Environment and Biometrics Department, Forestry Research Institute of Nigeria, Ibadan, Nigeria.*
[2]*Department of Statistics, Federal Polytechnic, Ile-Oluji, Nigeria*

*daveadedoyin@yahoo.com

Article Info

Received: 13/04/2021
Accepted: 17/09/2021

Keywords
Survival times, Censoring, Proportional Hazards, Likelihood estimation

Abstract

Survival observations are incomplete and often with ties consequently resulting in bias when the full likelihood estimation method is used and ties are not considered. This study compares the Exact, Efron and Breslow methods of the Cox proportional hazards (PH) model using survival times of breast cancer in–patients admitted to University of Ilorin Teaching Hospital, Kwara State. Three hundred patients were diagnosed with breast cancer. Of this figure, only ninety–seven (approximately 32.3%) patients experienced the event of interest (i.e. death) while only eighty–six were uncensored and used in this study. The log–cumulative hazards plot and Schoenfeld's residual test affirmed fulfilment of the PH assumption. Patients with cancer in the left breast were observed to have lower chance of survival compared with those with cancer in the right breast while female breast cancer patients were observed to have higher probability of survival compared to the male patients. Log–rank test showed no significant difference in gender disparity ($p = 0.153$) while a significant difference was observed for cancer location disparity ($p = 0.027$), at 5% alpha level. Similar estimated and standard error values were obtained for each variable using Breslow, Efron and Exact methods. Age and Cancer location were observed to be significant risk factors for breast cancer. The Exact, Breslow and Efron approximation methods for Cox PH regression model will produce approximately similar results when the PH assumption is fulfilled and moderate tie exist in the observations.

1. INTRODUCTION

Proportional hazards (PH) regression models are analytical techniques for analysing time–to–event measured observations. The branch of learning involving the analysis of such observations is popularly referred to as survival analysis. Survival analysis finds application in several varied fields such as medical or clinical studies, sociology, demography, environmental studies, psychology, engineering sciences etc. In medical sciences, for instance, survival analysis is used to assess the time to occurrence of events recovery of subjects placed on some form of treatments or time from diagnosis of an ailment, particularly when deadly, to the time of death. In engineering on the other hand, this study assists in reliability study to determine time to failure of a manufactured product (e.g. machine parts or electronics). A central theme for all disciplines applying this tool is that the observations involve time from, or to, the occurrence of an event of interest while the fundamental object of interest is to make reliable management decisions.

Survival analysis dataset however are characteristically non–normal and incomplete owing to censoring and truncation. According to [1], censoring is described as a loss of information on individuals or subjects due to failure of follow–up, withdrawal of subject(s) from study, study

termination when subjects have different dates of enrolment or death from competing risk. Generally, three types of censoring occur in practice: (i) right censoring: which occurs when subjects leave the study before the occurrence of the study of interest or the study ends before the event has occurred, (ii) left censoring: involving the occurrence of the event of interest before subjects were enrolled in the study or before observation time, and (iii) interval censoring: where the event of interest occur only within an interval. Since the full likelihood function was implemented to handle studies involving complete observations (i.e. no missing data), it therefore means that its application when there is censoring or truncation will introduce bias to the estimated regression parameters [2]. Consequently, the partial likelihood function has been developed to address this problem.

Cox's proportional hazards model, belonging to the semi–parametric model family, and accelerated failure time model, belonging to the parametric model family, are the most common proportional hazard tools used to date. Between the two however, the Cox model appears better preferred owing to its fewer assumptions which makes its implementation simpler and its ability to give robust parameter estimates once the observations it is used for fulfils the proportional hazards criterion/assumption and have no tie (i.e. no two survival time is the same) [3, 4, 5]. However, experience with real life survival observations are such that often present ties. Consequently, once the proportional hazards assumption is fulfilled, ties between the observations should be attached importance [6]. It was on this basis that the Exact [7] approximation method was developed to handle the problem of tie between time to event observations. This method however is computationally intensive and usually impractical when working with large datasets. The Breslow [8] and Efron [9] approximation methods later developed, which are computationally simpler compared to the Exact method, were then developed. [3] opined Efron method to be of choice on the basis of the better parameter estimates and fit statistics it provided and its relative faster computation time compared to the other two (Breslow and Exact methods). [10] on the other hand, in a study to assess the "validity and efficiency of approximation methods for tied survival time in Cox regression" observed that Efron approximation method performed far better for moderate or heavy ties than the Exact and Breslow approximations. A common feature of these two studies however is that they are based on simulated dataset. This study therefore compares the Exact, Efron and Breslow methods of estimation using real–life dataset with moderately tied failure times in a Cox proportional hazards model setting.

2. MATERIALS AND METHOD

2.1 Data

The dataset used for this study is a secondary data, originally sourced by [11] from the Cancer Registry department of the Admission and Discharge unit of University of Ilorin Teaching Hospital, Nigeria. In their paper titled, "Breast cancer patients in Nigeria: Data exploration approach," the distribution of the individuals included in the study was discussed using only descriptive tools and logistic regression model.

The dataset contains information on three hundred (300) in–patient breast cancer patients, comprising two hundred and seventy–five (275) women and twenty–five (25) men, who were observed over a five–year period (2011–2016). The patients were all treated as in–patients and were later discharged. From the three hundred observations, ninety–seven (97) were discharged dead while the remaining two hundred and three (203) patients were either discharged alive or lost to follow–up hence were censored out from this study. The variables examined in the dataset are:

i. Age: Age of patients in years;
ii. Sex: Gender of patients (0 = Male; 1 = Female);
iii. LOS: Survival time in days

iv. Location of Cancer: Breast cancer location at presentation (left breast, right breast and both breasts) and
v. Outcome: Event indicator (0 = Alive, 1 = Dead)

2.2 Methods

Cox Proportional Hazards (PH) Model

The Cox PH model is a semi–parametric model which is usually presented in terms of the hazards function. This model gives an expression for the hazard time t for an individual with a given set of explanatory variables. Consequently, let $\boldsymbol{x}' = (x_1, x_2, \dots, x_p)$ be a set of explanatory variables, called risk factors, accounting for the occurrence of an event of interest Y and let $\boldsymbol{\beta}$ denote a vector of parameters of size p. Also let C denote the censoring time and (T, δ) be a function defined on Y and C such that $T = \min(Y, C)$ and $\delta = I(Y < C)$, where

$$\delta = \begin{cases} 1 \; if \; Y < C \; (uncensored \; observation) \\ 0 \; if \; Y > C \; (censored \; observation) \end{cases}$$

is an indicator variable with values 1 *or* 0. Then, the Cox Proportional Hazard model is given in the as [12, 13, 14]:

$$h(t|\boldsymbol{x}) = h_o(t) \exp(\boldsymbol{x}'\boldsymbol{\beta}) \tag{1}$$

where $t = uncensored$ survival time;
$h(t|\boldsymbol{x}) = h$azard function at time t given the covariates **x**;
$h_o(t) =$ baseline hazard function; and
$\beta = vector \; of \; p \; covariates \; effects.$
The proportional hazards assumption for the model were investigated using Log –cumulative hazard plot and Schoenfeld (or partial) residual plot.

Cox PH Model's Parameter Estimation

Since survival analysis observations are characterized by loss of information, the parameters of the Cox PH model was estimated using Partial likelihood function. According to [12], [1], [3], [13] and [14], let n comprise the observed survival times in a study such that n is partitioned into k distinct uncensored survival times , ranked as $t_{(1)}, < t_{(2)} < \cdots < t_{(k)}$, and $n - k$ censored survival times**.** Also, let there be a risk set denoted $R(t_{(i)})$ associated with the survival times $t_{(i)}$. Then the probability that a particular failure is observed is presented as:

$$[f(t_i, \beta_i, x_i)]^{\delta_i} = [h(t_i, \beta_i, x_i)]^{\delta_i} [S(t_i, \beta_i, x_i)]^{\delta_i} \tag{2}$$

The likelihood for this function is:

$$L(\beta) = \prod_{i=1}^{n} \{[h(t_i, \beta_i, x_i)]^{\delta_i} [S(t_i, \beta_i, x_i)]\} \tag{3}$$

resulting in the form, by some simple substitution:

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_j\right)}{\sum_{l \, \epsilon \, R(t_{(i)})} \exp\left(\sum_{j=1}^{p} \beta_j x_j\right)} \right\}^{\delta_i}. \tag{4}$$

The partial log–likelihood, $l_p(\boldsymbol{\beta})$, for expression (4) is given as:

$$l_p(\boldsymbol{\beta}) = \mathbf{log}_e\left[\prod_{i=1}^{k}\frac{\delta_i \exp\left(\sum_{j=1}^{p}\beta_j x_{j(i)}\right)}{\sum_{l\,\epsilon\,R(t_{(i)})}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)}\right] \tag{5}$$

which can be written in the form:

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{k}\delta_i\left\{\sum_{j=1}^{p}\beta_j x_{j(i)} - \log_e\left[\sum_{l\,\epsilon\,R(t_{(i)})}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)\right]\right\} \tag{6}$$

The log–likelihood function presented on equation (6) holds when no tie exist in survival time of the observations. When ties exist however, the Exact method (i.e. equation 7) of the log–likelihood is preferred. According to [7], the partial likelihood when there are ties is:

$$l_p(\boldsymbol{\beta}) = \prod_{i=1}^{m}\frac{\exp\left(\sum_{u=1}^{r}\beta_u s_u\right)}{\sum_{P\epsilon Q_i}\prod_{k=1}^{d_i}\left[\left(\sum_{u=1}^{r}\exp(\boldsymbol{\beta_u x_{ul}})\right)\right]} \tag{7}$$

where P is one of the elements of Q_i, representing a set of d_i events occurring at t_i distinct ordered times $(i = 1,2,\dots,m)$, defined as $P = (P_i, P_i, \dots, P_{d_i})$ and set $D(t_i) = \{i_1, i_2, \dots, i_{d_i}\}$ representing a set of labels for failing observations at t_i.

When the number of ties is large, computation using the Exact method becomes a problem resulting to the use of Breslow's method of Cox Partial likelihood function (equation 8). This was achieved by summing up covariate related components for all subjects experiencing the event at a given time–point. Breslow's Cox Partial likelihood function is given as:

$$l_p(\boldsymbol{\beta}) = \prod_{i=1}^{m}\frac{\exp\left(\sum_{u=1}^{r}\beta_u s_u\right)}{\left\{\sum_{u\epsilon R(t_i)} exp\left(\sum_{u=1}^{r}\beta_u x_{ul}\right)\right\}^{d_i}} \tag{8}$$

Efron's method for the partial likelihood function for tied observations is however a further adjustment to the Breslow's method given in equation (8) when the ties are relatively large. It is given as:

$$l_p(\boldsymbol{\beta}) = \prod_{i=1}^{m}\frac{\exp(\boldsymbol{\beta' s_i})}{\prod_{k=1}^{d_i}\left[\sum_{u\epsilon R(t_{(i)}} exp(\boldsymbol{\beta' x_l}) - (k-1)d_i^{-1}\sum_{u\epsilon D(t_{(i)})}\exp(\boldsymbol{\beta' x_l})\right]} \tag{9}$$

However, when no ties exist in the datasets, all the three methods (Exact, Breslow and Efron) give the same result. When ties exist, Efron's and exact approximations have been documented to give relatively the same result.

3. RESULTS AND DISCUSSIONS

A summary of the dataset is presented on Table 1. Three hundred patients were diagnosed with breast cancer; twenty–five males (8.3%) and two hundred and seventy–five females (91.7%). Ninety–seven (32.3%) patients experienced the event of interest, here death, within this period and were uncensored while the remaining two hundred and three patients were censored from the study. The patients observed in this dataset were observed to either have cancer diagnosed in one of their breasts (left or right) or both breasts. So as to allow for comparison of survival when cancer is detected on each breast, patients who had cancer diagnosed in both breasts were exempted from the study, leaving eighty–six (28.7%) uncensored cancer patients for the analysis.

The age range (measured in years) of patients diagnosed with breast cancer in this study is (24, 96) while the median age is 50 years (Table 2).

Table 1: Summary for breast cancer patients at UITH (2011–2016)

Location of cancer	Gender	Event status		
		Censored	Uncensored	Total
Left breast	Male	4 (**2.9%**)	5 (**3.6%**)	9 (**6.4%**)
	Female	82 (**58.6%**)	49 (**35.0%**)	131 (**93.6%**)
Right breast	Male	10 (**7.4%**)	5 (**3.7%**)	15 (**11.1%**)
	Female	93 (**68.9%**)	27 (**20.0%**)	120 (**88.9%**)
Both breasts	Male	1 (**4.0%**)	0 (**0.0%**)	1 (**4.0%**)
	Female	13 (**52.0%**)	11 (**44.0%**)	24 (**96.0%**)
Total	Male	15 (**5.0%**)	10 (**3.3%**)	25 (**8.3%**)
	Female	188 (**62.7%**)	87 (**29.0%**)	275 (**91.7%**)

Table 2: Summary for Age of breast cancer patients (in years) at UITH (2011–2016)

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
24	40	50	50.34	60	96

Kaplan–Meier's survival plot for patients' survival probability partitioned in respect to the breast location where the cancer was diagnosed is presented on Figure 1. From the survival plot, patients who had cancer diagnosed in the right breast were observed to have a better chance of survival compared to those diagnosed with cancer in the left breast at 95% confidence interval.

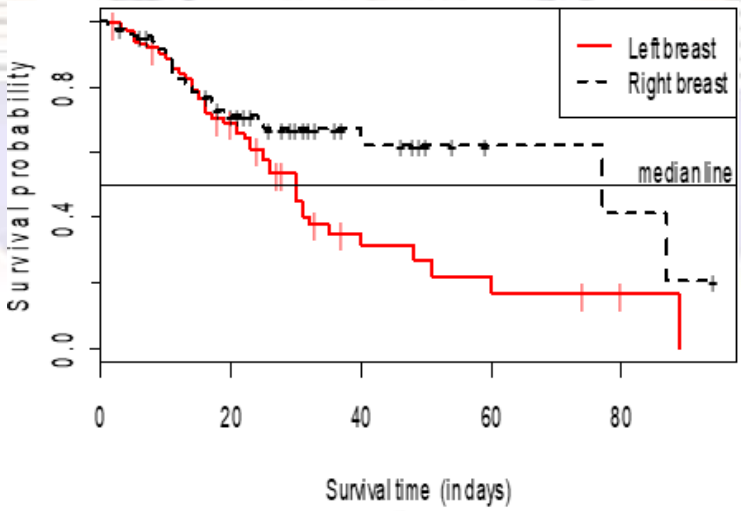

Figure 1: Kaplan –Meier survival plot for breast cancer survival by breast location (2011–2016)

The median survival time for patients diagnosed with cancer in the left breast was observed to be thirty (30) days while those diagnosed with cancer in the right breast had survival time of seventy–seven (77) days (Table 3).

Table 3: Median Survival time by cancer location

Variable	Number of patients	# Outcome	Median time	0.95 LCL	0.95 UCL
Left breast	140	54	30	25	40
Right breast	135	32	77	40	

Female breast cancer patients were also observed to have higher chance of survival compared with the male patients who had breast cancer (Figure 2). The median survival time for females with breast cancer is 40 days while their corresponding male counterpart is 24 days (Table 4).

Table 4: Median survival time by gender for breast cancer patients

Variable	Number of patients	# Outcome	Median time	0.95 LCL	0.95 UCL
Female	250	76	40	30	87
Male	25	10	24	16	

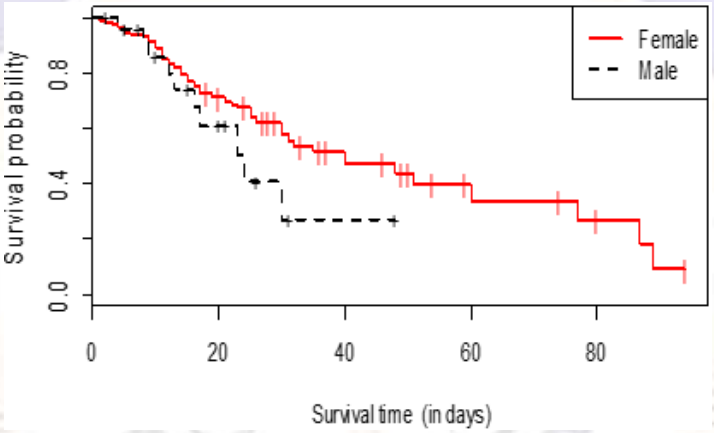

Figure 2: Kaplan–Meier survival plot for breast cancer survival by gender

The Log–Rank Test

The log–rank test was conducted to determine whether the observed differences on Figure 1 and Figure 2 were significant. From the test, the p–values for the gender and location of breast cancer variables were observed at the 95% confidence level to be 0.153 and 0.0271, respectively (Table 5, Table 6). These values signify that the difference in survival time observed on the Kaplan plot for the male and female patients (Figure 2) was not statistically significant while the difference in survival time for patients diagnosed with cancer in the left breast was statistically significantly different from the survival time of patients diagnosed with cancer in their right breast.

Table 5: Log–Rank test for breast cancer patients by Gender

Variable	Number	Obs. value	Exp. Value	$\frac{(Obs.-Exp.)^2}{Exp.}$	$\frac{(Obs.-Exp.)^2}{v}$
Female	250	76	79.46	0.151	2.04
Male	25	10	6.54	1.835	2.04

Chi-sq. = 2.0 ; df =1 ; p =0.153

Table 6: Log–Rank test for breast cancer patients by breast location

Variable	Number	Obs. Value	Exp. Value	$\frac{(Obs.-Exp.)^2}{Exp.}$	$\frac{(Obs.-Exp.)^2}{v}$
Left breast	140	54	43.9	2.33	4.88
Right breast	135	32	42.1	2.43	4.88

Chi-sq. = 4.9 ; df =1 ; p =0.0271

Log Cumulative Hazard Plots

Figure 3 and Figure 4 are graphical presentations of the log–cumulative hazards plot versus log survival time for the breast cancer patients observed on the gender variable and location of cancer in the breast, respectively. From the figures (that is, Figure 3 and Figure 4), the line plots for the male and female in the gender variable were observed to be approximately parallel (Figure 3) while plots for left breast cancer patients vs. right breast cancer patients were observed to be approximately parallel with each other (Figure 4). However, this may not give sufficient basis for concluding that the variables satisfy the proportional hazard assumption. Hence, a need for the Schoenfeld's partial residual test.

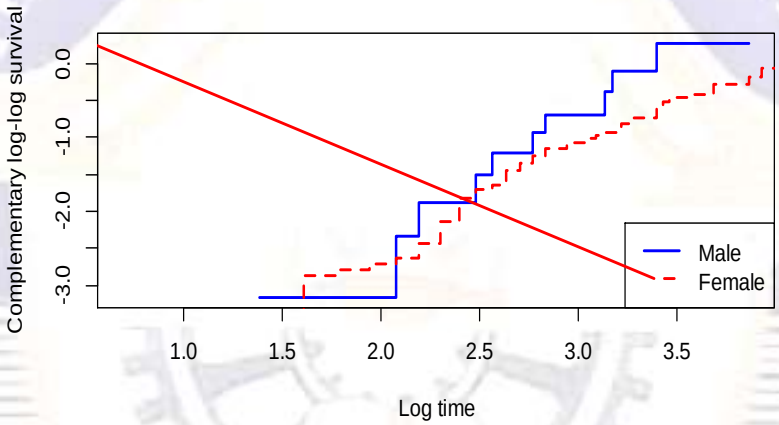

Figure 3: Log–cumulative hazards plot for breast cancer patients by gender

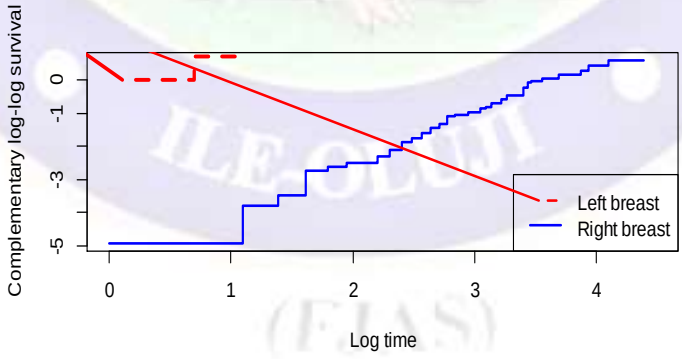

Figure 4: Log–cumulative hazards plot for breast cancer patients by breast location

Schoenfeld Partial Residual Tests

Results from the Schoenfeld's partial residual test (Table 7), showed a p–value that was greater than the alpha level for the age, gender and location of breast cancer variables at 95% confidence level. Hence, age, gender (with male gender as the baseline) and location of breast cancer (with patients with right breast cancer as baseline) variables satisfy the proportional hazard assumption.

Table 7: Schoenfeld's residual test for study variables

Variable	Rho	chis-sq.	p-value
Age	0.0236	0.0575	0.8105
Gender: Male	0.0958	0.7635	0.3822
Location of Cancer: Right breast	-0.1838	2.9318	0.0869

Parameter estimation for Cox Proportional Hazards Model

Table 8 presents a summary of results for the Breslow, Efron and Exact Cox proportional hazards model for the tied survival times. Each estimation method showed that age and location of breast cancer at diagnosis significantly affected hazard rate of patients while Gender was a non–significant factor. The coefficient values of the age and gender variables were also observed to be positive while the coefficient for the location of cancer was negative for all the Cox proportional hazard models adopted.

From Table 8, using Breslow's Cox PH estimates, an increase in patient's age is observed to significantly increase the hazard rate of patients suffering from breast cancer. In other words, a unit increase in patients age is observed to increase the hazard function by an approximate constant factor of 1.0256 (i.e. $e^{0.02563}$=1.02596). In the same vein, the gender variable was observed not to have a significant effect in explaining the hazard of the patients. However, the positive value of the coefficient (i.e. β_{gender} = 0.3673) and the male gender as the reference point implies that male breast cancer patients have higher hazard rate, with a constant factor of 1.4438, compared to the female breast cancer patients.

Furthermore, the location of breast cancer in patients have significant effect on hazard rate. However, a negative coefficient estimate for the "breast cancer location" variable and right breasts as the reference point indicates that patients with breast cancer on the right side of their breast had a significantly lower hazard rate, with a constant value of 0.6756, compared with patients who have breast cancer on their left breast.

The estimates of the parameters of the Cox PH model derived using Breslow, Exact and Efron partial log–likelihood methods (Table 9) showed that Breslow and Efron methods to have the least standard errors (≈ 0.0075), thought relatively same as for the Exact method if approximated in three decimal places. Consequently, no ties existed in the survival observations used in this study and each of the three methods of estimation performed equally well.

Table 8: Parameter estimates for tied survival times using Cox proportional hazards model

Method	Variables	Coef.	Exp(Coef.)	SE(Coef.)	Z	pr(>|z|)	Remark at 99% CI
Breslow's Cox Partial Likelihood method	Age	0.02563	1.025964	0.00745	3.440	0.00058	Significant
	Gender	0.36727	1.443796	0.3439	1.068	0.28549	Not significant
	Loacation of Cancer	-0.39215	0.675601	0.22666	-1.730	0.08361	Significant
Efron's Cox Partial	Age	0.02601	1.026351	0.007478	3.484	0.00049	Significant
	Gender	0.37584	1.45622	0.34389	1.093	0.27443	Not significant

Likelhood Function	Loacation of Cancer	-0.39829	0.671467	0.22663	-1.757	0.07883	Significant
Exact Cox Partial Likelihood Function	Age	0.02631	1.02666	0.00759	3.467	0.00053	Significant
	Gender	0.37626	1.45682	0.34964	1.076	0.28187	Not significant
	Loacation of Cancer	-0.40064	0.66989	0.22896	-1.750	0.08014	Significant

Baseline variable for Gender = Male; Baseline for Location of Cancer = Right breast

Table 9: Parameter estimates for the Parametric and Cox Proportional Hazard Models

Variable	MODEL		
	Breslow Cox	**Efron Cox**	**Exact Cox**
Age	0.0256 (**0.0075**)	0.0260 (**0.0075**)	0.0263 (**0.0076**)
Gender (Male)	0.3673 (**0.3439**)	0.3758 (**0.3439**)	0.3763 (**0.3496**)
Cancer location (Right)	-0.3922 (**0.2267**)	-0.3983 (**0.2266**)	-0.4006 (**0.2290**)

Entries in each cell represents (parameter (SE (parameter)

4. CONCLUSIONS

In this study, a Cox proportional hazards model was developed for breast cancer survival observations obtained from University of Ilorin Teaching Hospital. Patients with breast cancer in the right breast was observed to have higher chance of survival compared with those with breast cancer in the left breast. Also, female breast cancer patients tend to have better chance of survival compared to their male counterparts. All three estimation methods used (Efron, Breslow and Exact) agreed that Age and Location of breast cancer significantly increased the risk of death associated with breast cancer while gender was not significantly attributed to death from breast cancer.

With these findings, patients with breast cancer in the left breast will require more timely attention compared to those with cancer in the right breast. Any of the three estimation methods (Efron, Breslow and Exact) may be used for estimating the Cox proportional hazards model involving survival variables when moderate tie exist as approximately equal results will be produced.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Lee, E.T. and Wang, J.W. (2003). Statistical methods for survival data analysis (3rd ed.). U.S.A.: John Wiley and Sons, Inc.

[2] Teshnizi, S.H. and Ayatollahi, S.M.T. (2017). Comparison of Cox proportional and parametric models: application for assessment of survival of pediatric cases in acute leukemia in Southern Iran. Asian Pac. J. Cancer Prev, 18(4):981–85.

[3] Borucka, J. (2014). Methods for handling tied events in the Cox proportional hazard model. *Studia Oeconomica Posnaniensia,* 2(2):91–106.

[4] Harrell, F.E. Jr. (2001). Regression modelling strategies: with application to linear models, logistic regression and survival analysis. USA: Springer Science + Business Media Inc.

[5] Kleinbaum, D.G. and Klein, M. (2012). Survival analysis: A self–learning text (3rd ed.). USA: Springer Science and Business media.
[6] Cox, D.R. (1972). Regression models and Life–tables. *J. Royal Stat Soc.* (Series B), 34(2):187–220.
[7] Kalbfleisch, J.D. and Prentice, R.L. (2002). The statistical analysis of failure time data (2nd ed.). USA: Wiley & Sons, Inc.
[8] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics,* 30:89–99.
[9] Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association,* 75(359):557–565. http://dx.doi.org/10.1080/01621459.1977.10480613
[10] Hertz–Picciotto, I. and Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics,* 53:1151–1156.
[11] Oguntunde, P.E., Adejumo, A.O. and Olkagbue, H.I. (2017). "Breast cancer patients in Nigeria: Data exploration approach". Data in Brief, 15:47–57. http://dx.doi.org/10.1016/j.dib.2017.08.038
[12] Jackson, R.J. (2015). Some statistical methods for the analysis of survival data in cancer clinical trials (Ph. D thesis). University of Liverpool, England.
[13] Collet, D. (2003). Modelling survival data in medical research (2nd ed.). USA: Chapman & Hall/CRC.
[14] Liu, X. (2012). Survival analysis models and applications. United Kingdom: John Wiley & Sons, Inc.
[15] Nikulin, M.S., Balakrishnan, N., Mesbah, M and Limnois, N. (2004). Parametric and semiparametric models with applications to reliability, survival analysis and quality of life. USA: Springer Science + Business Media Inc.