sciendo

DOI: 10.2478/bile-2024-0001

Biometrical Letters
Vol. 61 (2024), No. 1, 1-15

Survival Analysis of Covariates Influencing Breast Cancer Treatment: A Case Study of North Eastern Nigeria

Shadrach Yusuf, E. Nkiruka Okeke, Kazeem E. Lasisi

Department of Mathematics and Statistics Federal University Wukari. Taraba State Nigeria,
e:mail: shadrachyusuf86@gmail.com

SUMMARY

This study builds on previous research indicating that breast cancer survival time is influenced by several underlying factors. The study covered a period of 10 years from January 2012 to December 2022, and 140 cases were considered within the study cohort. The study considered breast cancer patients from North East Nigeria. The methodologies used are Cox hazard proportional regression and Kaplan–Meier analysis. The mean patient survival time is 592.2 days, with an average age of 44.61 years, average number of children of 5, and mean weight difference of 1.95 kg while on treatment. Kaplan-Meier analysis and the log rank test were used to investigate how the various covariates affect survival time, and it was found that age and family history have significant effects on the survival time in the studied population. The p-value of 0.04 for radiotherapy indicates statistical significance, in contrast to other treatment options such as surgery (p-value 0.7), targeted therapy (p-value 0.7), and chemotherapy (p-value 0.6). Residual diagnostic analysis with a component for assessment of Variance Inflation Factors (VIF) was used to detect multicollinearity among the independent variables. A total of 60 events (deaths) occurred within the study period with a concordance value of 0.73, which indicates a moderate level of agreement. This implies that the model's predictions align reasonably well with the observed outcomes.

Key words: breast cancer, survival time, multicollinearity, Cox hazard, Kaplan–Meier analysis

1. Introduction

Background of study

Patients suffering from various diseases may have variable chances of survival. The likelihood of survival from certain diseases is dependent on the patient's body defence mechanism (Halid, 2019), access to and affordability of medical facilities, taking of appropriate medications, etc. This research will consider the

rate (chances) of survival for breast cancer patients. Survival may be described as the ability of a human being to withstand a persistent trait which may occur as a result of interaction with the environment. Survival rate is a term used to describe the probability of a patient remaining alive for a stipulated period of time, and survival time specifies the length of period taken for failure to occur (Daniel & Cross, 2013). Cancer is a generic term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumours and neoplasm. Malignant cases involve cells with a high ability to spread to other parts of the body through either the blood or lymph system (Machin et al., 2007).

One defining feature of cancer is the rapid formation of abnormal cells which grow beyond their usual boundaries and can invade adjoining parts of the body and spread to other organs. The latter process is referred to as metastasis, and leads to the development of secondary malignant growths at a distance from the primary site of the cancer. Metastases are a major cause of death from cancer diseases. The most common cancers, with their respective numbers of reported cases, are lung cancer (2.09 million cases), breast cancer (2.09 million), colorectal cancer (1.80 million), prostate cancer (1.28 million), non-melanoma skin cancer (1.04 million), and stomach cancer (1.03 million), according to figures from the World Health Organization (WHO, 2018).

Cancer is the second leading cause of death globally after ischemic heart disease, and was responsible for an estimated 9.6 million deaths in 2018. Globally about 1 in 6 deaths is due to cancer. Approximately 70% of cancer deaths occur in low- and middle-income countries. In 2017, the Nigeria National Cancer Control Plan was announced, covering the period from January 2018 to December 2022, with an estimated cost of over 97 billion naira. It was reported that cancer is responsible for 72,000 deaths yearly in Nigeria, with an estimated 102,000 new cases annually. Around one-third of deaths from cancer are due to the five leading behavioural and dietary risks: high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco, and alcohol (WHO 2018).

Survival analysis denotes a collection of statistical techniques used to describe and quantify time-to-event data. In survival analysis we use the term

failure to denote the occurrence of the event of interest, even though the event may actually be a success, such as recovery from therapy (Lee and Wang, 2003). The term survival time specifies the length of time taken for failure to occur. Studies related to survival include clinical trials, prospective cohort studies, retrospective cohort studies, and retrospective correlation studies (Rosner, 2010).

Adamu et al. (2019) applied the Cox proportional hazard regression model to a survival analysis of cancer patients in North Eastern Nigeria from 2004 to 2017. Oladuti and Olopha (2022) performed a comparison of survival estimation methods in the analysis of breast cancer survival. In this study we considered more covariates that may influence breast cancer survival time, such as treatment options, level at presentation, and socio-economic factors including age, with a view to developing a model that will estimate patients' survival time.

2. Research methodology

2.1. Data sources

The data for this study were obtained from the Federal Teaching Hospital, Gombe. Breast cancer patients' files were retrieved from the hospital records and from oncology departments. The following variables were considered: date of diagnosis, sex, age, marital status, number of children, state of residence, occupation, stage at presentation according to the Tumour Nodes Metastases (TNM) system, treatment option (chemotherapy, surgery, targeted therapy, hormonal treatment, radiography), histological type, date of last visit, family history, substance taking, anthropometries, and patient's status (alive or death). Patients and their relatives were contacted by telephone for purposes of data collection.

2.2. The Kaplan–Meier estimator

In the following, methods for estimating the survivor, hazard and cumulative hazard functions associated with a set of survival data will be presented. The log rank test is a method for comparing two or more groups of survival times (Machin

et al., 2006). When none of the survival times in a survival data set is censored, the survivor function can be estimated by the empirical survivor function, given by:

$$S(t) = \frac{Individuals\ in\ the\ cohort\ with\ survival\ times}{Total\ individuals\ in\ the\ cohort}; f\ ort \geq 0$$

$$S(t) = 1 - G(t)$$

where $G(t)$ is the empirical cumulative distribution function defined by

$$G(t) = \frac{Individuals\ alive\ in\ the\ cohort at\ time\ t}{Total\ individuals\ in\ the\ cohort}$$

2.3. Cox regression

Consider two groups of patients, all diagnosed with the same disease but receiving different treatments. One group receives standard treatment, while the other receives a new treatment. Each group has an associated hazard function at time (t), denoted as $h_S(t)$ $h_S(t)$for patients on standard treatment and $h_N(t)$ $h_N(t)$for those on the new treatment. The proportional hazard model is defined as:

$$h_N(t) = \psi h_S(t) \quad (1)$$

where ψ is the relative hazard or hazard ratio. Depending on whether $\psi < 1$or $\psi > 1$there is a lower or higher risk of death associated with the new treatment compared with the standard treatment (Folorunso et al., 2015). We can express equation (1) as:

$$h_i(t) = e^{\beta x_i} h_0(t) \quad (2)$$

The symbols in (2) are as follows: $h_i(t)$is the hazard function for patient *i*; X_i is either 0 or 1, indicating the patient's treatment; $\beta = ln(\psi)$; $h_0(t)$is the baseline hazard function.

Now, assume we have *p* explanatory variables $X_1, X_2, X_3, \ldots, X_p$with values $x_1, x_2, x_3, \ldots, x_p$at the start of the study. Define the vector x containing the values of the explanatory variables as:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix} \tag{3}$$

In addition, we define $h_0(t)$ as the baseline hazard function, which corresponds to $x = 0$ (all explanatory variable values are zero). The hazard function for patient i is then given by

$$h_i(t) = \psi(x_i)h_0(t) \tag{4}$$

where $\psi(x_i)$is the relative hazard function associated with the explanatory variable values of patient i, and $x_i$is the column vector given by

$$x_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \\ \vdots \\ x_{pi} \end{bmatrix} \tag{5}$$

The relationship in (5) is also known as the general proportional hazard model. We can rewrite $\psi(x_i)$as

$$\psi(x_i) = e^{\beta^T x_i} \tag{6}$$

where$\beta^T x_i$is defined as:

$$\beta^T x_i = \sum_{j=1}^{p} \beta_j x_{ji} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_p x_{pi} \tag{7}$$

The $\beta_i s$are the coefficients of the explanatory variables $x_1, x_2, x_3, \ldots, x_p$in the model, and $\beta^T x_i$ is called the linear component of the model.

$\beta^T$is the transpose of a column vector containing the coefficients of the p explanatory variables:

$$\beta^T = [\beta_1 \quad \beta_2 \quad \beta_3 \quad \ldots \quad \beta_p] \tag{8}$$

With the above terminology, the expression for the Cox proportional hazard model is given by (Nelima, 2013):

$$h_i(t) = e^{\beta^T x_i} h_0(t).$$

We will provide a comprehensive presentation of the results derived from the Cox regression analysis and offer in-depth interpretations of these findings. Cox regression analysis is a powerful statistical tool that allows us to understand the impact of various covariates on survival outcomes. By examining the coefficients, p-values, and hazard ratios associated with each covariate, we can gain a deeper understanding of the factors that significantly influence survival within the study population (Folorunso et al., 2015).

2.4. Simulation

In this section, we will explore the simulation of survival times and their relationship to various factors such as the number of children, age, and weight. The data were analysed using Cox regression. We started by simulating survival data with the assumption that they follow the Weibull distribution, the age and weight difference with the assumption that they follow the normal distribution, and the number of children with the assumption that it follows the uniform distribution (Venkataramanaiah et al., 2015).

To simulate random variables from a given probability distribution, we used the Inverse Transform Method. This method involves generating uniform random variables and then transforming them into variables with the desired distribution. We used the built-in functions rnorm, rweibull and runif of the R package for the various simulations (Sheppard, 2010).

3. Analysis and discussion of results

3.1. Descriptive summary of data

This section provides a comprehensive and detailed description of the covariates considered in this study. In the following subsections, we will expound on the various factors and variables that were taken into account as covariates, shedding light on their significance and relevance to the present research. The elucidation of these covariates is essential for a thorough understanding of how

they may influence the study's outcomes, making it an integral part of the research methodology (Peet & John, 2019).

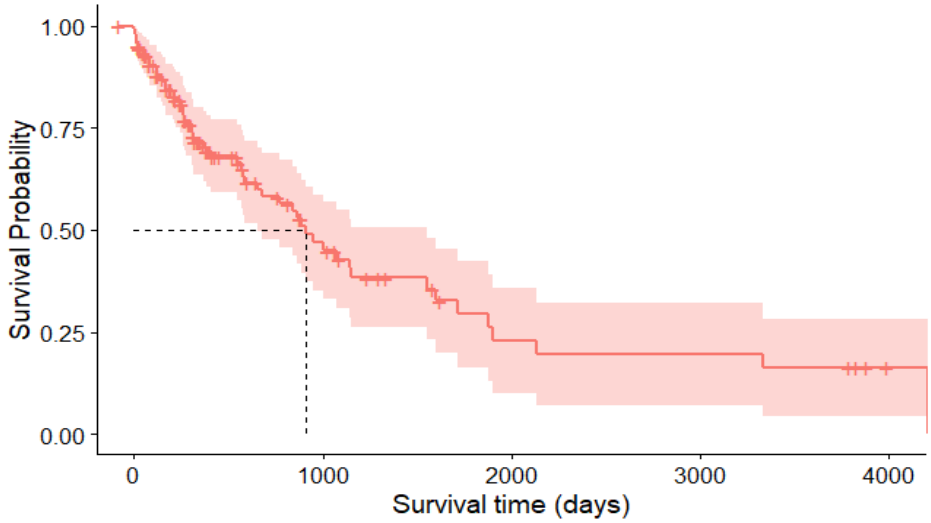

Figure 1. Survival time plot for all cases

In Figure 1, the survival time for the entire cohort is graphically represented. The y-axis of the plot depicts the survival probabilities, while the x-axis corresponds to the survival time of the cohort, measured in days. In accordance with the principles of Kaplan–Meier (KM) analysis, the plot indicates that the total survival probability for the entire cohort at time 0 (the starting point) is equal to unity (1), signifying that all participants were alive at the beginning of the study. This probability gradually decreases as the number of days increases (t = 1, 2, 3,…, n), reflecting the diminishing likelihood of survival over time. Notably, the plot also revealed that the median survival time for the cohort was less than 1000 days, indicating that approximately half of the participants experienced the primary event of interest (in this case death) within that time frame. This information is pivotal for understanding the overall survival dynamics within the study cohort.

Table 1 presents a summary of the log rank test for each of the covariates in the KM analysis.

Table 1. Summary of Log Rank Test

Variable	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V	Chi-Square	P-value
Age							
0–25	2	1	0.214	2.877	2.91	6.5	0.0900
26–50	107	49	43.773	0.624	2.35		
51–75	30	9	15.394	2.656	3.64		
75–85	2	1	0.618	0.236	0.24		
Sex							
Male	5	1	3.66	1.931	2.1	2.1	0.1
Female	136	59	56.34	0.125	2.1		
Marital status							
Divorced	7	2	1.73	0.0432	0.0432	2.9	0.4
Married	111	50	47.34	0.1492	0.7128		
Single	8	1	4.36	2.5930	2.8280		
Widow	15	7	6.57	0.0286	0.0325		
Family history							
No	50	13	23.8	4.91	8.66	8.7	0.003
Yes	91	47	36.2	3.23	8.66		
Residence type							
Rural	52	22	22.9	0.0342	0.0583	0.1	0.8
Urban	89	38	37.1	0.0211	0.0583		
Stage at diagnosis							
1.0	3	0	2.29	2.28523	2.44209	3.7	0.4
2.0	5	2	1.21	0.51850	0.53832		
3.0	32	12	14.14	0.32302	0.42702		
4.0	20	9	9.26	0.00723	0.00871		
5.0	81	37	33.11	0.45686	1.03720		
Comorbidities							
No	11	3	3.27	0.02189	0.0235	0.0	0.9
Yes	130	57	56.73	0.00126	0.0235		
Surgery							
No	91	37	38.3	0.0419	0.12	0.1	0.7
Yes	50	23	21.7	0.0737	0.12		
Radiotherapy							
No	137	60	56.2	0.258	4.14	4.1	0.04
Yes	4	0	3.8	3.805	4.14		
Targeted therapy							
No	91	37	38.3	0.0419	0.12	0.1	0.7
Yes	50	23	21.7	0.0737	0.12		
Chemotherapy							
No	7	3	2.16	0.3256	0.344	0.3	0.6
Yes	134	57	57.84	0.0122	0.344		
Weight difference							
Increase	40	17	22.5	1.362	2.47	2.5	0.1
Decrease	101	43	37.5	0.819	2.47		
Number of children							
0–3	14	5	5.9	0.1366	0.1530	0.3	0.9
4–6	74	34	32.3	0.0872	0.2029		
7–9	44	19	19.8	0.0309	0.0487		

Table 2. Cox regression summary

Covariates	Coefficient	Hazard Ratio	Lower CI	Upper CI	P-value	Remark
Age category: (25,50)	-4.020	0.018	-6.687	-1.353	0.003	S
(50,75)	-4.727	0.009	-7.670	-1.785	0.002	S
(75,85)	-5.639	0.004	-9.460	-1.818	0.004	S
Sex: M	-5.943	0.003	-9.020	-2.866	0.000	S
Marital status: Married	1.291	3.638	-0.641	3.223	0.190	NS
Single	-2.119	0.120	-5.028	0.790	0.153	NS
Widow	1.136	3.114	-0.809	3.081	0.252	NS
NOC	-0.051	0.950	-0.239	0.136	0.591	NS
Occupation: C/servant	-1.584	0.205	-3.131	-0.038	0.045	S
Farmer	3.513	33.550	1.096	5.930	0.004	S
H/wife	-1.906	0.149	-3.277	-0.536	0.006	S
Others	-2.166	0.115	-3.974	-0.357	0.019	S
Family history: Yes	1.177	3.244	0.334	2.020	0.006	S
Residence type: Urban	0.534	1.705	-0.102	1.170	0.100	NS
Substance taking: Yes	-0.862	0.422	-3.176	1.452	0.465	NS
Stage at diagnosis	-0.197	0.821	-0.735	0.341	0.473	NS
Comorbidities: Yes	0.343	1.408	-0.959	1.644	0.606	NS
Chemotherapy: Yes	-0.155	0.857	-1.448	1.138	0.814	NS

NS = not significant; S = significant, NOC = number of children

The probability values associated with various covariates and their respective categories in the context of survival analysis provide important insights into the potential influences on survival outcomes. Among these, the p-value of 0.003 for family history stands out as statistically significant, suggesting that the presence or absence of a family history of a particular condition significantly affects survival within the study population. Conversely, several covariates, such as sex, marital status, residence type, stage at diagnosis, comorbidities, surgery, targeted therapy, chemotherapy, weight difference, and number of children, have p-values

higher than 5%, indicating that these variables do not exhibit significant impact on survival time. The results further revealed that its usage may significantly influence survival outcome.

The Cox regression analysis results presented in Table 2 provide valuable insights into the relationship between various covariates and their impact on survival outcomes, with reference to the associated coefficients, hazard ratios, confidence intervals, and p-values.

3.2. Significant covariates

Within the different age categories, a significant association with survival emerges. Specifically, individuals aged 25–50, 50–75, and 75–85 exhibit hazard ratios of 0.018, 0.009, and 0.004, respectively. These hazard ratios are significantly below 1, signifying that as age progresses within these categories, the risk of the event (death) significantly diminishes. Furthermore, the low p-values of 0.003, 0.002, and 0.004 bolster the significance of these findings. Male sex (M), as indicated by a coefficient of –5.943, is clearly linked to a reduced risk of the event, with a hazard ratio of 0.003. The exceptionally low p-value of 0.000 underscores the statistical significance of this effect. Moreover, the occupation of farmer is associated with a notably elevated hazard ratio of 33.550, signalling a substantial increase in the risk of the event. The low p-value of 0.004 reaffirms the statistical significance of this relationship. Lastly, a positive family history significantly heightens the risk of the event, as indicated by a hazard ratio of 3.244, supported by a low p-value of 0.006, confirming the statistical significance of this association.

3.3. Non-significant covariates

Several covariates, including marital status (married, single, widow), number of children (NOC), residence type (urban, rural), substance taking, stage at diagnosis, comorbidities, and chemotherapy, do not exhibit significant effects on survival outcomes. Their coefficients either are not statistically significant or

have hazard ratios close to 1, suggesting that they do not substantially influence survival within the studied population.

Table 3. Variance inflation factor for the covariates

Covariates	GVIF	Df	GVIF^(1/(2*Df))
Age category	4.977	3	1.307
Sex	2.386	1	1.545
Marital status	9.738	3	1.461
No. of children	1.744	1	1.321
Occupation	16.098	4	1.415
Family history	1.722	1	1.312
Residence type	1.381	1	1.175
Substance taking	1.212	1	1.101
Stage at diagnosis	1.212	1	1.101
Comorbidities	1.202	1	1.096
Chemotherapy	1.202	1	1.096
Surgery	1.643	1	1.282
Weight difference	1.269	1	1.126

It is observed from the results that some covariates have higher GVIF values than others, indicating potentially higher multicollinearity.

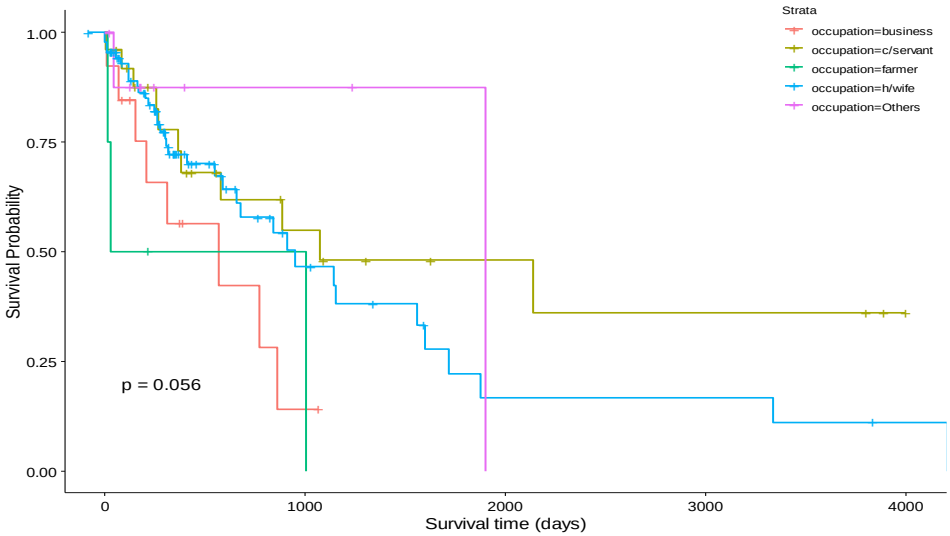

Figure 2. KM curves with respect to the occupation covariate

Figure 2 shows that the occupation covariate is a significant factor which influences treatment outcomes. Those working in farming do not exceed 1000 days of treatment; those working in business die within 1000 days of treatment. Civil servants and housewives survive up to 4000 days of treatment.

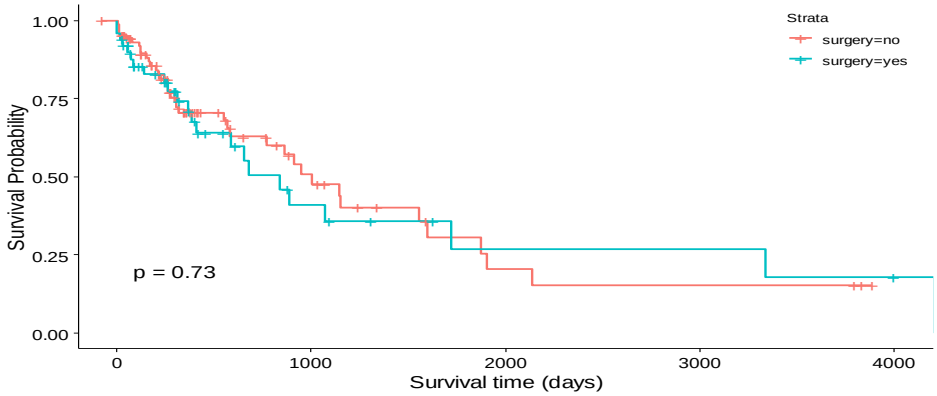

Figure 3. KM curves with respect to the surgery covariate

Figures 3 shows that surgery is not a significant factor, because in most cases the affected area regenerates and the cancer spreads again.

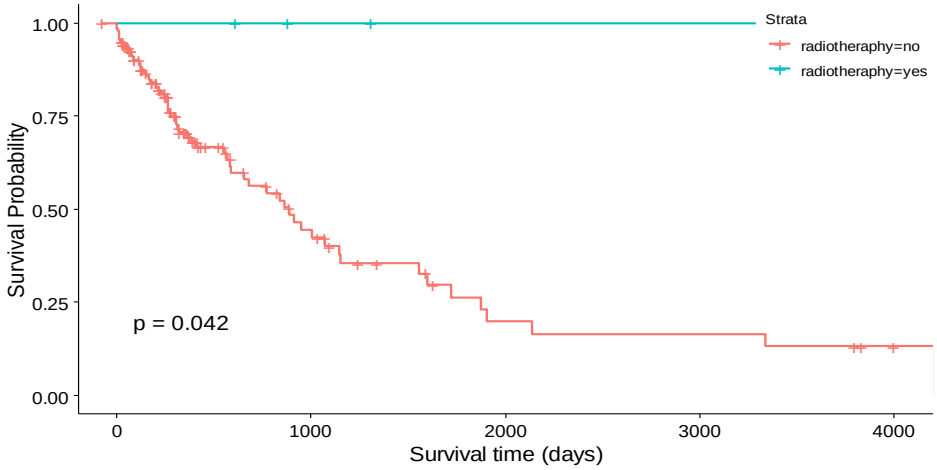

Figure 4. KM curves with respect to the radiotherapy covariate

Figure 4 shows that radiotherapy is a significant covariate, because it destroys cancer cells using a strong beam of light, preventing them from regenerating.

4. Simulation

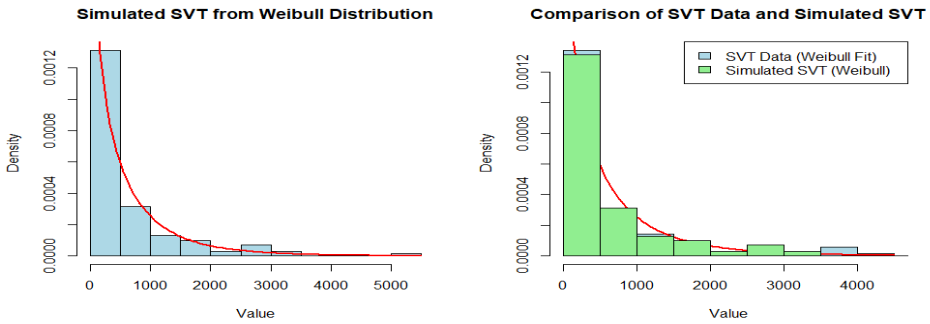

Figure 5. Simulated SVT

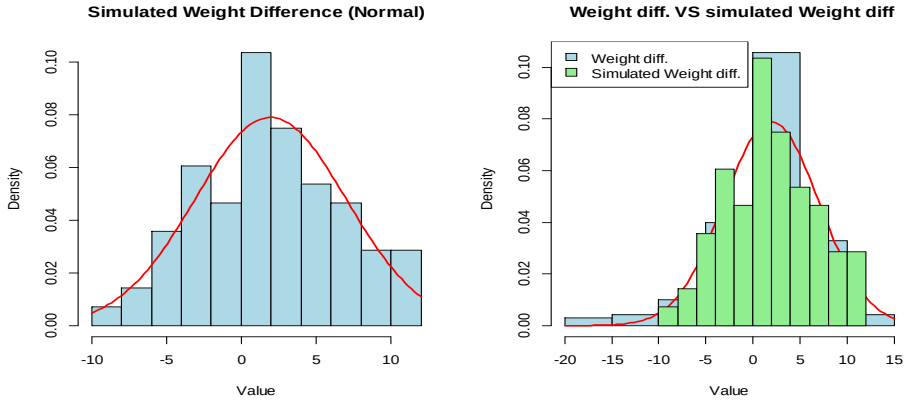

Figure 6. Simulated weight difference

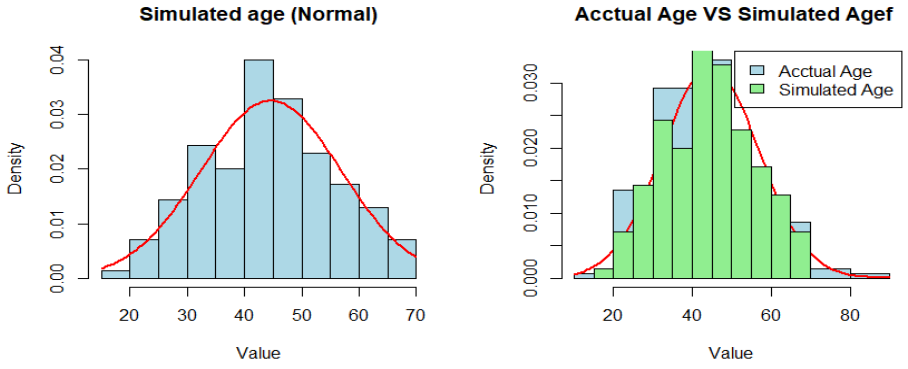

Figure 7. Simulated age

Figures 5, 6, and 7 offer visual representation of the original variables and provide a side-by-side comparison between the simulated and original variables. These graphical presentations reveal that the differences between the simulated and original data are minimal. The plots effectively show the similarity between the original and simulated variables, emphasizing the accuracy of the simulation process in reproducing the key characteristics of the original data.

5. Conclusion

The study was conducted to determine the effect of certain covariates on the survival time of breast cancer patients. The Kaplan–Meier analysis (Figure 1) highlights that the entire cohort's survival probability decreases over time, with the median survival time being less than 1000 days, indicating that approximately half of the participants experienced the primary event (death) within this time frame. The assessment of Variance Inflation Factors (VIF) reveals low multicollinearity, ensuring stable coefficient estimates.

Gender-specific health programmes: The strong association between male sex and a reduced risk of the event should encourage further investigation into the factors contributing to this disparity. This may lead to the development of gender-specific health programmes or interventions aimed at reducing risks for females, particularly in relation to the primary event of interest.

Occupational health initiatives: The notably elevated hazard ratio for the "Farmer" occupation category suggests a need for specific health initiatives or support systems for individuals in this occupation. Understanding the occupational factors contributing to this increased risk may inform targeted interventions.

Family history assessment: The significant positive impact of family history on the risk of the event highlights the importance of collecting and assessing family medical histories during clinical evaluations. Clinicians should pay close attention to this factor when determining treatment plans and risk assessments. The result indicating that radiotherapy significantly influenced survival outcomes

suggests that healthcare providers should carefully consider its inclusion in treatment plans.

REFERENCES

Adamu P.I., Adamu M.O., Okagbue H.I., Opoola L., Bishop S.A. (2019): Survival Analysis of Cancer Patients in North Eastern Nigeria from 2004-2017 A Kaplan–Meier Method. Journal of Medical Sciences. 7(4): 643-650

Daniel W.W., Cross C.L. (2013): Biostatistics A Foundation for Analysis in the Health Science. [10th Edition] John Wiley & Sons. Ltd: 750-773

Federal Ministry of Health, Nigeria. (2017): National Cancer Control Plan

Folorunso S.O., Chukwu A.U., Oluwasola T.A. (2015): The Cox Regression Model With Application To Predict Admission Lifetime Or Female Genital Cancer.

Halid O.Y., Ogunboyo O.F., Adebayo K. (2019): A Survey of Survival Patterns of Some Recidivists in Nigerian Prisons European Journal of Scientific Research

Harris M., Taylor G. (2003): Medical Statistics Made Easy. Martin Dunitz Taylor and Francis Group New York: 57-60

Limenih Y.A., Workie D.L (2019): Survival Analysis of Time to Cure on Multi-drugs resistance Tuberculosis Patients in Amhara region, Ethiopia BMC Public Health 19: 165

Machin D., Campbell M.J., Walters S.J. (2007): Medical Statistics A Textbook for the Health Sciences. [4th Edition]. John Wiley & Sons. Ltd: 181-198

Machin D., Cheung Y.B., Parmar M.K. (2006): Survival Analysis A Practical Approach. John Wiley & Sons. Ltd South Gate, England: 23-155

Nelima K.E. (2013): Estimating the Survival of Patients with Cancer of the Cervix at Kenyatta National Hospital in Nairobi, Kenya (Master's thesis, Jomo Kenyatta University of Agriculture and Technology, Kenya 2013).

Lee E.T., Wang J.W. (2003): Statistical Methods for Survival Data Analysis. John Wiley & Sons. Ltd Canada [3th Edition]: 64-196

Oladuti M.O., Olopha P.O. (2022): Comparison of Survival Estimation Methods in the Analysis of Breast Cancer Data. Asian Pacific Journal of Cancer Care 7: 601-605

Peat J., Barton B. (2005): Medical Statistics A Guide to Data Analysis and Critical Appraisal. Blackwell Publishing: 202-266

Peet, M., Broggio J. (2019): The Impact of Updating Cancer Survival Methodologies for National Estimates. Office for National Statistics and Public Health, England

Rosner B. Fundamentals of Biostatistics BROOKS/COLE CENGAGE learning (2010): [7th Edition]: 783-787

Sheppard A.J. (2010): Breast Cancer Survival in Ontario First Nations Women Understanding the Determinants (PhD Dissertation, Institute of Medical Sciences University of Toronto).

Venkataramanaiah A.B. (2015): Statistical Applications of Survival Data Analysis for Breast Cancer Data

World Health Organisation (2018): Cancer National Plan