Available online at www.worldnewsnaturalsciences.com

World News of Natural Sciences

An International Scientific Journal

WNOFNS 58 (2025) 163-182 EISSN 2543-5426

Systematic assessment of Cox proportional hazards, exponential, log-normal survival models in time to event using breast cancer data

Uchechukwu Kalu[1,a], Lamidi Kehinde Rasheed[1,b], Okechukwu Uzoma Iheme[2], Akorede Yussuf Toheeb[1], Ugonna Uchechi Iheme[3]

[1] Departement of Mathematics and Statistics, Kwara State University, Malete, Kwara State, Nigeria

[2] Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystem,
South Botanical Garden, Chinese Academy of Science, Guangzhou, China

[3] Department of Dental Technology, Federal University of Technology,
PMB 1526, Owerri, Imo State, Nigeria

[a,b]E-mail address: kaluuchechukwu45@gmail , rasheedlamidi44@gmail.com

ABSTRACT

This study systematically evaluates the performance of Cox proportional hazards, exponential, and Log-normal survival models using a dataset of 230 breast cancer patients. Descriptive statistics reveal a predominance of female patients (96%) and various cancer stages, with the majority at Stage II (41%). The Kaplan-Meier curve illustrates a gradual decline in overall survival probability over 35 months, dropping to approximately 50% by 20 months. Significant differences in survival probabilities are observed based on smoking status ($p = 0.006$) and occupation ($p = 0.001$), while no significant differences are detected across cancer stages ($p = 0.5$) or treatment types ($p = 0.1$). The Cox model indicates that smoking status and specific occupations significantly affect hazard ratios, while immunotherapy shows a significant reduction in hazard (HR = 0.609, $p = 0.018$). The proportional hazards assumption remains largely intact across the covariates in the Cox model. The comparison of survival models using AIC and BIC values shows that the Log-Normal model performs best, with the lowest AIC (1255.282) and BIC (1302.461), indicating a better fit while accounting for model complexity. The Cox Proportional Hazards model ranks second with an AIC of 1385.218 and a BIC of 1424.698. The Exponential model, with the highest AIC (1402.989) and BIC (1464.875), fits the data least effectively. Overall, the Log-Normal model provides the best balance between accuracy and simplicity in this analysis.

(Received 28 October 2024; Accepted 29 November 2024; Date of Publication 01 December 2024)

***Keywords*:** Cox Proportional Hazards, Exponential Model, Log-Normal Model, Breast Cancer Data, Time-To-Event Analysis

1. INTRODUCTION

The analysis of time-to-event data is a crucial component of survival analysis, especially in the context of clinical and epidemiological research on diseases like cancer (Smith et al., 2022). Breast cancer, in particular, stands out as one of the most common malignancies, affecting millions of women worldwide, Due to the complex and varied pathways of disease progression and mortality, advanced statistical techniques are essential to capture the diverse survival patterns seen in patients (Bray et. al., 2018).

Key methodologies that have gained prominence in predicting and analyzing survival times in breast cancer patients include the Cox proportional hazards model, the exponential model, and the log-normal survival model (Schober P, Vetter TR. 2018; Smith et al., 2022; Nassif et al., 2022).

The assessment of survival models is crucial in cancer research, as it provides insights into the timing of clinical events, which can significantly impact treatment decisions and patient prognostication (Giordano et.al., 2018). The Cox proportional hazards model, introduced by Sir David Cox in 1972, is particularly notable for its semi-parametric nature, allowing for evaluating the effect of various covariates on the hazard function while making fewer assumptions about the underlying survival distribution. Its flexibility and interpretability have made it the cornerstone of survival analysis in clinical studies, including those that focus on breast cancer, the ability to accommodate both categorical and continuous predictors renders the Cox model adept for analyzing diverse patient characteristics, treatment modalities, and demographic factors that may influence survival time (Schober P, Vetter TR. 2018).

On the other hand, parametric models such as the exponential and log-normal survival models offer particular advantages in specific situations. The exponential model, known for its simplicity, assumes a constant hazard rate over time (Schober P, Vetter TR. 2018). While this straightforward approach can be useful for initial analyses, its assumptions may fall short when the hazard rate exhibits more complex patterns.

In contrast, the log-normal model provides greater flexibility, allowing the hazard rate to vary over time, and is better suited for capturing the nuanced survival patterns seen in breast cancer patients (Schober P, Vetter TR. 2018). The selection of an appropriate survival model plays a critical role in shaping the conclusions drawn from the data, underscoring the importance of comparing these models to enhance our understanding of breast cancer survival dynamics (Barker & O'Connell, 2021).

Breast cancer survival data often encompasses a range of covariates, such as age, tumor size, nodal involvement, and treatment modalities, that may influence patient outcomes. The interaction of these factors creates a complex landscape that necessitates robust statistical modeling.

Previous studies have demonstrated the importance of tailoring statistical approaches to account for these covariates, leading to better predictive accuracy and improved clinical decision-making. As such, the need for thorough assessments of various survival models arises, particularly in the context of emerging data and research techniques.

In this evaluation, it is crucial to examine how each model's assumptions correspond to the dataset's characteristics (Miller & Hodge, 2020). For example, the Cox model assumes proportional hazards, meaning the hazard ratio between any two individuals remains consistent over time. This assumption can be evaluated through diagnostic tools like Schoenfeld residuals, which help detect potential violations. Conversely, the exponential model's assumption of a constant hazard rate may lead to overly simplistic conclusions when applied to data with clear time-varying hazards. When the log-normal model is more suitable, the dataset may show distributions that challenge the assumptions of simpler models, requiring advanced techniques such as maximum likelihood estimation to handle these complexities (Wang & Wei, 2020).

Additionally, the move towards a more comprehensive approach to patient care highlights the importance of including patient-reported outcomes and quality-of-life indicators in survival analyses (Smith et al., 2022). By incorporating these aspects into statistical models, researchers can provide insights into survival rates and the lived experiences of breast cancer patients. However, this expansion of data introduces greater complexity, underscoring the demand for survival models capable of effectively integrating these diverse elements of patient experience (Gomez & Lammers, 2022; Ryosuke Fujii, 2023).

Breast cancer research is experiencing rapid progress, with the use of large datasets and advanced analytical methods becoming more prevalent. Machine learning and artificial intelligence are increasingly employed to analyze and model survival data, offering the ability to uncover complex patterns that might be missed by traditional approaches (Wang & Wei, 2020). Despite this, fundamental statistical methods like Cox regression and parametric survival models remain essential, and comparing their effectiveness across different datasets continues to be vital for advancing the field (Wang & Wei, 2020).

A thorough systematic Assessment of the Cox proportional hazard model, exponential model, and log-normal survival model using breast cancer data will advance the discussion on the best analytical approaches for survival analysis, this analysis seeks to highlight the strengths and weaknesses of each method concerning specific datasets, ultimately improving the accuracy of time-to-event outcome modeling. Moreover, the findings will provide valuable insights for clinicians, researchers, and policymakers in selecting the most suitable statistical tools for analyzing breast cancer survival, contributing to better patient care strategies (Kim & Kim, 2021).

This paper is structured as follows: In Section 2.1, we present the Non-parametric models. Section 2.1.1 presents the Kaplan-Meier curve which is a visual representation showing the probability of surviving in a given length of time. Section 2.1.2 outlines the log-rank test which tends to compare the survival experience of two or more groups of individuals. Section 2.2 presents the semi-parametric model. Section 2.2.1 presents the Cox proportional hazards model. Section 2.3 presents Parametric models.

Section 2.3.1 presents the Exponential Model while 2.3.2 presents the Log Normal Model. Section 3.1 presents the descriptive statistics of the data. Section 3.2 presents results from the Kaplan-Meier curve and log-rank tests for the equality of survival functions (survival probabilities). Section 3.3 presents results from the Cox proportional hazards model. Section 3.4 presents results from the parametric Model (Exponential and Log-Normal Model), Section 3.5 will discuss results from the Cox pH, exponential model, and log-normal model Section 3.6 will evaluate the models used in the study of survival analysis using model comparison section 4 presents our closing remarks on the study.

2. MATERIALS AND METHODS

This study has appropriately used the Non-parametric Kaplan-Meier and log-rank test to predict the survival curve from time-to-event data and to assess the survival experience of two or more groups of individuals respectively, as well as using the Semi-parametric(Cox proportional hazard model) for analyzing individual or grouped survival times, and also the parametric model (Exponential and Log-normal model) to demonstrate how a priori an individual will survive with survival and many explanatory variables in a demonstration form. On the other hand, the log-normal model will tend to accommodate variation in the hazard over time, making it more flexible for representing survival patterns where the risk of events changes with time instead of following a constant frequency like that assumed in the Exponential model.

2. 1. Non-parametric approach

Non-parametric models are statistical models used in survival analysis. They are flexible and can accommodate complex consequences of covariates on the cure rate (Ghosh, 2021). A non-parametric approach is not restricted by assumptions concerning the nature of the population from which a sample is drawn. This is opposed to parametric statistics, for which a problem is restricted a priori by assumptions concerning the specific distribution of the population (such as the normal distribution) and parameters (such as the mean or variance). Nonparametric statistics is based on either not assuming a particular distribution or having a distribution specified but with the distribution's parameters not specified in advance (though a parameter may be generated by the data, such as the median). The non-parametric approach can be used for descriptive statistics or statistical inference. Non-parametric approaches are often used when the assumptions of parametric tests are violated.

2. 1. 1. Kaplan Meier

The Kaplan-Meier estimator is defined as the probability of surviving in a given time while considering time in many small intervals. The visual representation of this function is usually called the Kaplan-Meier curve, and it shows what the probability of an event (for example, survival) is at a certain time interval. If the sample size is large enough, the curve should approach the true survival function for the population under investigation (Jiang et al., 2023) the Kaplan-Meier estimator is mathematically represented as follows.

Given a set of survival times with $t_1, t_2, \dots \dots \dots \dots \dots \dots, t_n$ corresponding censoring indicators δ_i (where δ_i = 1 if the event was observed and δ_i = 0 if the observation was censored), the Kaplan-Meier estimator for the survival function is defined as:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{1}$$

- d_i is the number of events (e.g., deaths) that occurred at time t_i
- n_i is the number of individuals at risk just before time t_i,
- The product is taken over all distinct event time t_i up to time t.

2. 1. 2. The log-rank test

The log-rank test is a statistical method used to compare the survival experience of two or more groups of individuals (Schober P, Vetter TR. 2018). It is a nonparametric test and appropriate to use when the test is based on the times of events, such as deaths, and is used to test the null hypothesis that there is no difference between the populations in the probability of an event at any time point, the analysis is based on the observed and expected number of events in each group at each observed event time, the log-rank test is most likely to detect a difference between groups when the risk of an event is consistently greater for one group than another, the log-rank test is widely used in medical research to compare the survival distributions of two or more independent group, it is often used to compare the survival of patients to a reference survival curve that typically represents the expected survival under the standard of care, the log-rank test is most commonly used statistical test for comparing the survival distributions of two or more groups.

$$X^2 = \frac{(O_1-E_1)^2}{E_I} + \frac{(O_2-E_2)^2}{E_2} + \frac{(O_3-E_3)^2}{E_3} \dots\dots\dots. + \frac{(O_K-E_K)^2}{E_K} \qquad (2)$$

where k is the number of groups being compared, O_K is the observed number of events in the Kth group over time, and E_K is the expected number of events in the Kth group over time.

2. 2. Semi-parametric approach

The semi-parametric models are without any survival time distribution assumption, but other assumptions remain (e.g. the relationship between survival time and the covariate or the proportionality of hazards) (Carroll, 2021). A semi-parametric model is intermediate between parametric and nonparametric models and contains finite-dimensional and infinite-dimensional parameters, Semi-parametric models are advantageous because they strike a balance between the flexibility of non-parametric models and the structure provided by parametric models. They are often used when the assumptions of purely parametric or purely non-parametric models do not hold. Researchers use these models to explore the relationship between predictor variables and survival outcomes while allowing for complex and non-linear effects.

2. 2. 1. Cox proportional hazards model

As stated by (Medhat et al., 2015), the Cox regression model is a statistical theory of counting processes that unifies and extends nonparametric censored survival analysis. The approach integrates the benefits of nonparametric and parametric approaches to statistical inferences. The Cox proportional hazards regression model relates covariates to the hazard function as follows:

$$h(t/x) = \ h_o(t)c\left(B^i x\right) \qquad (3)$$

where: $h_o(t)$ is called the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the model are zero $B^i = B_1 + B_2 + \cdots\cdots\cdots\cdots$ $+B_P$ is a parameter vector of regression coefficients of X= $X_1 + X_2 + \cdots\cdots\cdots\cdots +$, is the value of the vector of explanatory variables for a particular individual, and c· is a fixed, known scalar function is the value of the vector of explanatory variables for a particular individual, and c (.) is a fixed, known scalar function.

This is a semi-parametric model where the baseline hazard $h_o(t)$ is estimated non-parametrically, while the covariate effect is constrained by the parametric representation $c\left(B^i x\right)$, where c (.) takes an exponential form:

$$c\left(B^i x\right) = e^{(B^i x)} = e^{\sum_{j=1}^{p} B_i x_{ji}} \tag{4}$$

Which assures that the hazard is non-negative and assumes that covariate effects on the hazard are multiplicative. Therefore,

$$h(t/x) = \ h_o(t)c\left(B^i x\right) = h_o(t)e^{(B^i x)} = h_o(t)e^{\sum_{j=1}^{p} B_i x_{ji}} \tag{5}$$

The Cox model is called a proportional hazards model since the ratio of hazard rates of two individuals with covariate values x_1 and x_2, at time t is

$$\frac{h(t/x_1)}{h(t/x_2)} = \frac{h_o(t)e^{(B^i x_1)}}{h_o(t)e^{(B^i x_2)}} = \frac{e^{(B^i x_1)}}{e^{(B^i x_2)}} = e^{(B^i(x_1 - x_2))} \tag{6}$$

The hazard ratio is time-independent as the ratio does not depend on t. Since the hazard function at t given covariate x is

$$h(t/x) = \ h_o(t)e^{(B^i x)} \tag{7}$$

The Cumulative Hazard Function

$$H(t/x) = \ \int_0^t h(t/x)du = \int_0^t h(u)\, e^{(B^i x)}du = \ H_O(t)e^{(B^i x)} \tag{8}$$

Survival Function

$$S(t/x) = \ e^{-[H_O(t)e^{\left(B^i x\right)}]} \tag{9}$$

Probability Density Function

$$f(t/x) = h_o(t)e^{(B^i x)}e^{-[H_O(t)e^{\left(B^i x\right)}]} \tag{10}$$

2. 3. Parametric approach

Parametric models are a type of statistical technique used in survival analysis to model the relationship between the survival of an individual and several explanatory variables, (Carroll, 2021), In parametric survival analysis, all parts of the model are specified, including the hazard function and the effect of any covariates, The distribution of the outcome, which is the time to event, is specified in terms of unknown parameters, The parametric models used in this study are Exponential and Log-normal models.

2. 3. 1. Exponential model

The Exponential Distribution is a fundamental parametric model used extensively in survival analysis, reliability engineering, and queuing theory. It is distinguished by its simplicity and its constant hazard rate, making it a key tool for modeling time-to-event data.

Hazard Function: The hazard function for the Exponential Distribution is given by:

$$h(t) = \lambda \tag{11}$$

where λ is the rate parameter. This indicates that the hazard rate is constant over time, which implies that the likelihood of an event occurring is uniform.

Survival Function: The survival function, representing the probability of surviving beyond a time t, is expressed as:

$$\mathrm{S(t)} = e^{-\lambda \mathrm{t}} \tag{12}$$

This function decays exponentially over time, reflecting the constant hazard rate.

Cumulative Distribution Function: The cumulative distribution function, which gives the probability that an event has occurred by time t, is:

$$\mathrm{F(t)} = 1 - e^{-\lambda \mathrm{t}} \tag{13}$$

This function increases over time and approaches 1 as *t* grows larger.

2. 3. 2. Log-normal model

The log-normal model is a popular subgroup of survival analysis models used when the event times are not symmetrically distributed. It posits that the survival time of each individual follows a log-normal distribution, and hence is ideal for analyzing time-to-event data, where we are interested in knowing when an event will take place (length scales) within some specified period.

Survival Function: The survival function S(t) is the probability that an individual survives past time t. It is defined as follows: for a log-normal distribution, it is given:

$$S(t) = 1 - \phi\left(\frac{log(t) - \mu}{\sigma}\right) \tag{14}$$

where:

- ϕ is the cumulative distribution function (CDF) of the standard normal distribution
- μ is the location parameter (mean of the log-transformed survival time)
- σ is the scale parameter (standard deviation of the log-transformed survival time)
- t is the survival time.

Probability Density Function (PDF): The probability density function for the log-normal distribution is given as

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp(-\frac{(log(t)-\mu)^2}{2\sigma^2} \tag{15}$$

Hazard Function: The hazard function h(t) which is the instantaneous failure rate at time t, can be written as:

$$h(t) = \frac{\boldsymbol{f(t)}}{\boldsymbol{S(t)}} \tag{16}$$

This shows that the hazard rate for the log-normal distribution is non-monotonic, meaning that it increases to a peak and then decreases over time.

Cumulative hazard function: The cumulative hazard function H(t) is the integral of the hazard function over time and is expressed as:

$$H(t)= -log(S(t)) \tag{17}$$

3. RESULTS AND DISCUSSION

3. 1. Descriptive statistics for the data

Table 1. Gender distribution of respondents

Gender	Frequency	Percentage
Female	218	96%
Male	12	4%
Total	230	100%

From Table 1, it was discovered that 218 patients were female with a corresponding frequency of 96% while 12 patients were male with a corresponding frequency of 4%.

Table 2. Stages distribution of respondents

Stages	Frequency	Percentage
Stage I	60	26%
Stage II	95	41%
Stage III	45	20%
Stage IV	30	13%
Total	230	100%

From Table 2, it was discovered that patients at stage I of breast cancer were 60 with a corresponding percentage of 26% of the entire population, while patients at stage II of breast cancer were 95 with a corresponding percentage of 41% of entire the population while patients at stage III of breast cancer were 45 with a corresponding percentage of 20% of entire the population and lastly patients at stage IV of breast cancer were 30 with a corresponding percentage of 13% of entire the population.

3. 2. Results from Kaplan-Meier curve and log-rank tests for equality of survival functions (survival probabilities)

The log-rank test is a statistical test used to compare the survival distributions of two or more groups in a survival analysis and determine if there is a significant difference in survival times (time to an event of interest, such as death or failure) between different groups, the log-rank test calculates a test statistic based on the observed and expected number of events in each group, considering the time to event or censoring. The test statistic follows a chi-squared (χ^2) distribution, and the significance level of the test can be used to determine whether the survival curves are significantly different.

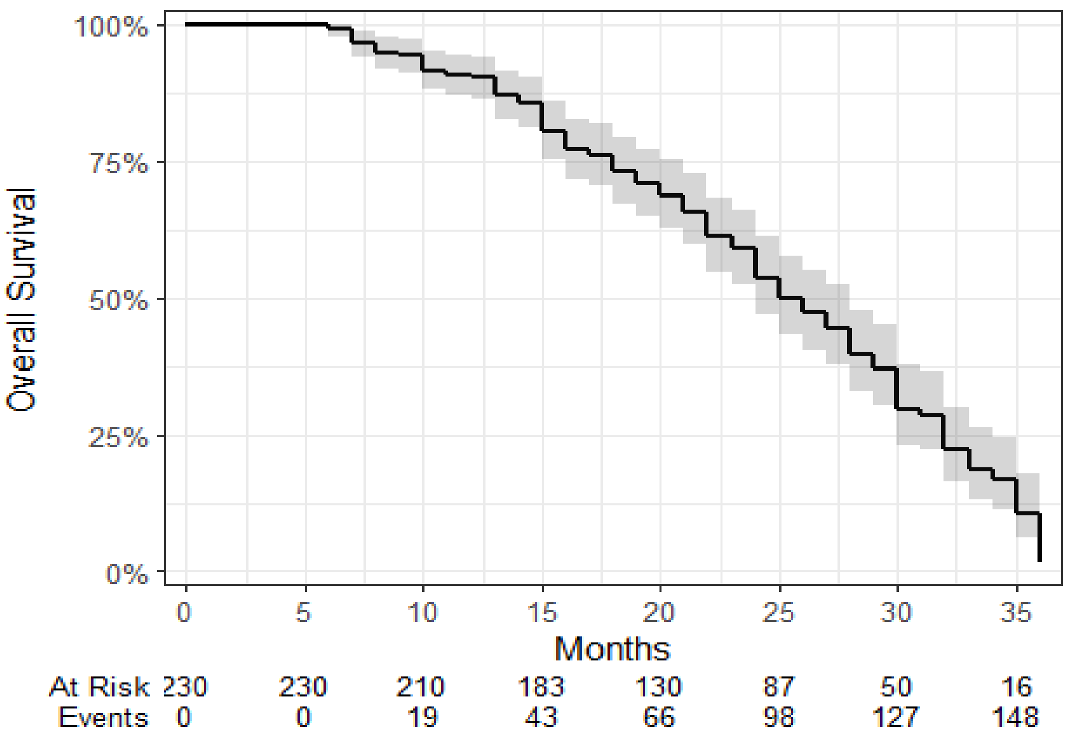

Figure 1. Kaplan-Meier curve on the overall survival

Figure 1 presents the Kaplan-Meier curve for the overall survival probability shows a gradual decrease in survival probability over 35 months. At the start, all 230 participants have a 100% survival probability, which slowly declines as events such as deaths or other outcomes occur. By approximately 18-20 months, the survival probability falls to around 50%, indicating that half of the participants have faced the event by this point. The decline is steady, with no sudden changes in survival rates over time. By the end of the study at 35 months, the survival probability nears 0%, suggesting that nearly all individuals have encountered the event. Additionally, the confidence intervals expand as time progresses, highlighting the growing uncertainty in survival estimates as fewer participants remain at risk toward the end of the study.

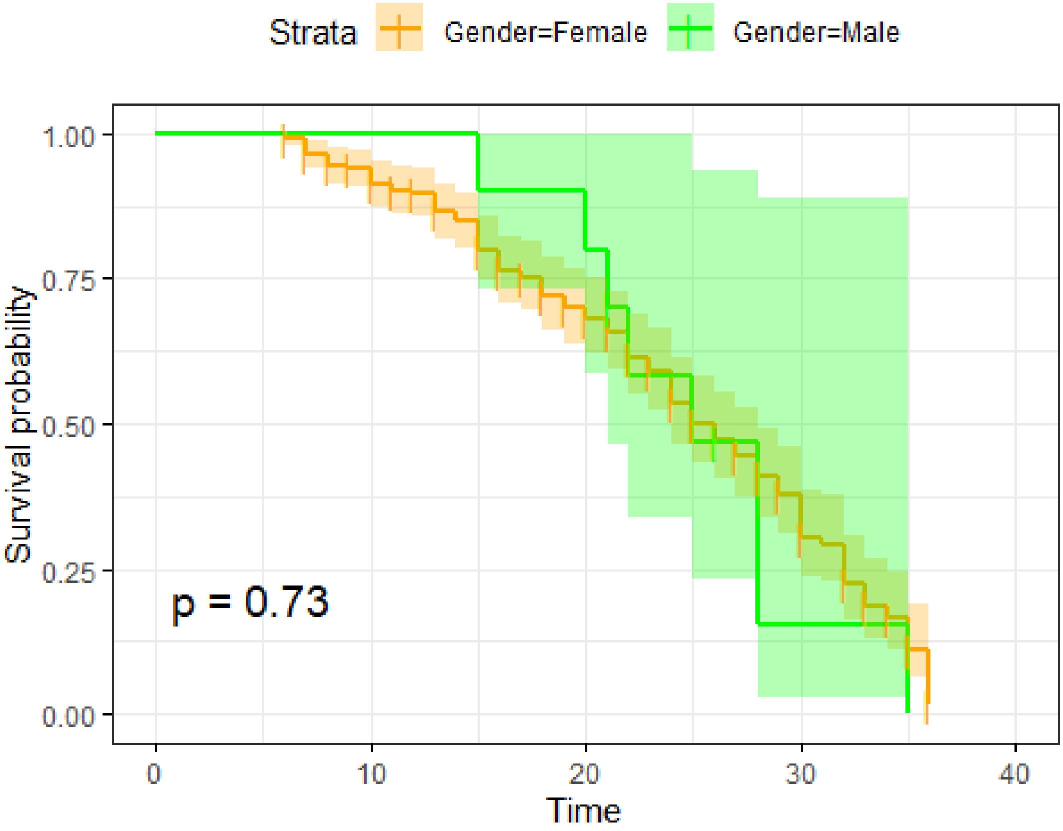

Figure 2. Kaplan Meier curve on gender.

Table 3. Results of smoking status of breast cancer patients

Smoking Status	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
Non-Smokers	118	74	77.3	0.137	0.308
Smokers	112	80	76.7	0.138	0.308

Chi-square = 0.3, P-value = 0.58

H_0 : There are no significant differences in survival probabilities between the two groups (Non-Smokers and Smokers)

H_i : There are significant differences in survival probabilities between the two groups (Non-Smokers and Smokers)

Figure 2 presents the Kaplan-Meier plot that compares survival probabilities between two groups: males (green) and females (orange). Both groups show a gradual decline in survival over time, with slight differences in survival probability throughout the study period. The p-value of 0.73 indicates no statistically significant difference between the survival curves of males and females. Confidence intervals, represented by the shaded areas, overlap considerably, further supporting the conclusion that gender does not significantly affect survival in this dataset. Both genders appear to have a similar median survival time, and any minor differences observed are likely due to random variation rather than a true effect.

Table 3 displays the log-rank test of 0.3 which indicates a low association level between the two groups (Non-Smokers and Smokers) regarding their survival probabilities, the p-value =0.58 $> \propto = 0.01$ suggests no statistically significant difference in survival probabilities between Non-Smokers and Smokers.

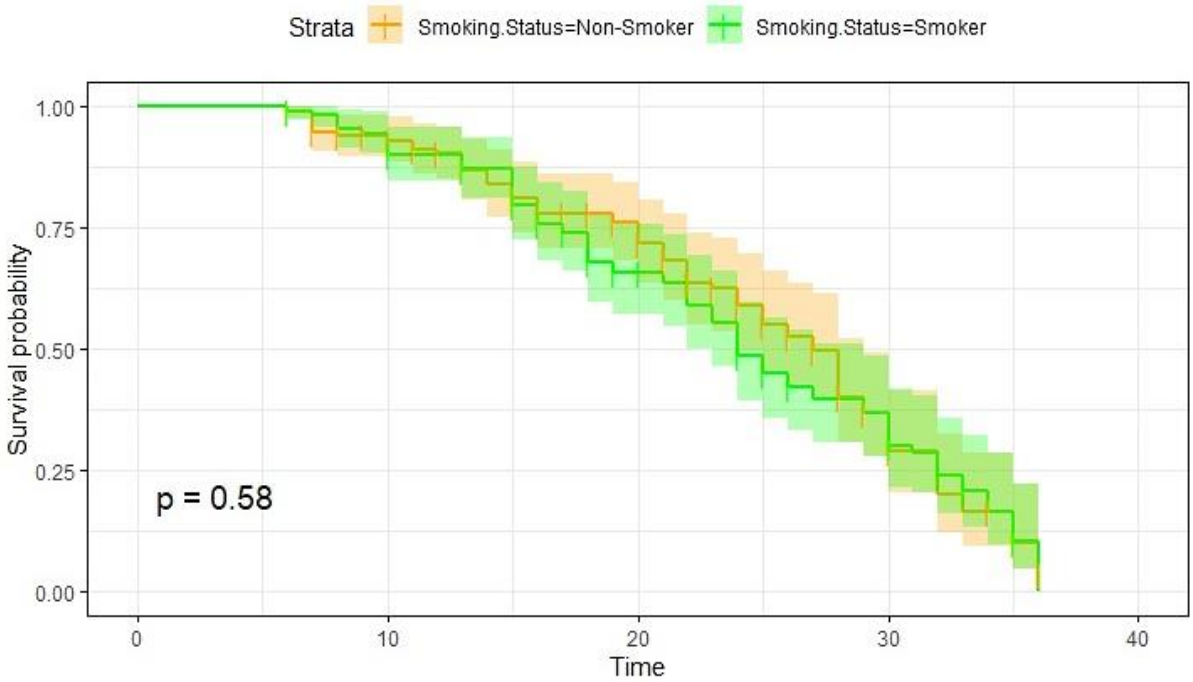

Figure 3. Kaplan Meier curve on smoking status

Figure 3 presents the survival curves showing that smokers tend to have lower survival probabilities over time compared to non-smokers.

Table 4. Results on occupation of breast cancer patients.

Occupation	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
1	170	114	110.060	0.1410	0.5622
2	22	17	14.537	0.4173	0.5201

3	7	5	6.206	0.2342	0.2705
4	11	6	6.372	0.0217	0.0276
5	2	2	0.208	15.4709	15.8703
6	18	10	16.618	2.6357	3.3979

Chi-square = 19.9, P-value = 0.001

H_0 : There is no significant difference in the survival probabilities for the six occupational categories of breast cancer patients

H_i : There is a significant difference in the survival probabilities for the six occupational categories of breast cancer patients

Table 4 reveals the log-rank test of 19.9 indicating a strong association between the occupational categories of breast cancer patients and their survival probabilities and the p-value = 0.001 < $\propto = 0.01$ suggesting a statistically significant difference in survival probabilities among the six occupational categories.

Table 5. Results on stages of breast cancer patients

Stages	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
I	61	38	43.7	0.7427	1.1751
II	64	44	44.7	0.0118	0.0187
III	60	47	40.0	1.2175	1.8387
IV	45	25	25.6	0.0121	0.0161

Chi-square = 2. 2, P-value = 0.5

H_0 : There is no significant difference in the survival probabilities for the four stages of breast cancer.

H_i : There is a significant difference in the survival probabilities for the four stages of breast cancer.

Table 5 reveals the log-rank test of 2.2 indicating a weak association between the stages of breast cancer and their survival probabilities and p-value = 0.53 $> \propto= 0.01$ this suggesting that there is no statistically significant difference in survival probabilities among the four stages of breast cancer and at such he survival probabilities for the four stages of breast cancer patients are the same across the group.

This Kaplan-Meier curve in Figure 4 suggests that while there are differences in survival probabilities among breast cancer stages, these differences are not statistically significant based on the given p-value.

Therefore, while Stage I patients tend to have better outcomes, the differences may not be strong enough to draw firm conclusions about the impact of cancer stage on survival.

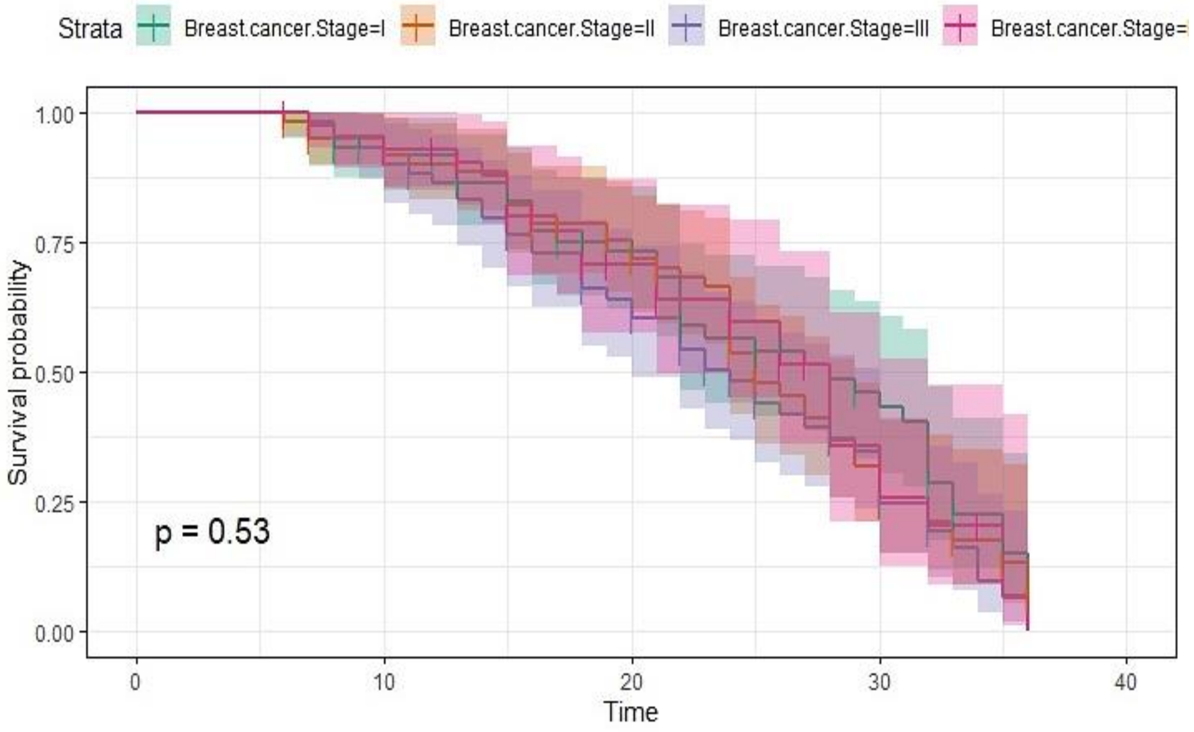

Figure 4. Kaplan-Meier curve on breast cancer stages

Table 6. Results of treatment types of breast cancer patients.

Treatment Type	N	Observed	Expected	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
Chemotherapy	77	53	46.6	0.890	1.413
Immunotherapy	73	48	59.5	2.239	4.216
Surgery	80	53	47.9	0.545	0.878

Chi-square = 4.2, P-value = 0.12
H_0 : There are no significant differences in survival probabilities between Treatment Type
H_i : There are significant differences in survival probabilities between Treatment Type

Table 6 shows that the log-rank test is 4.2 suggests a moderate association between treatment type and survival probabilities, and the p-value = $0.12 > \propto = 0.01$ indicates that there is enough evidence not to reject the null hypothesis H_0 , and conclude that the survival probabilities between treatment types of breast cancer patients are **the same** at a 1% level of significance.

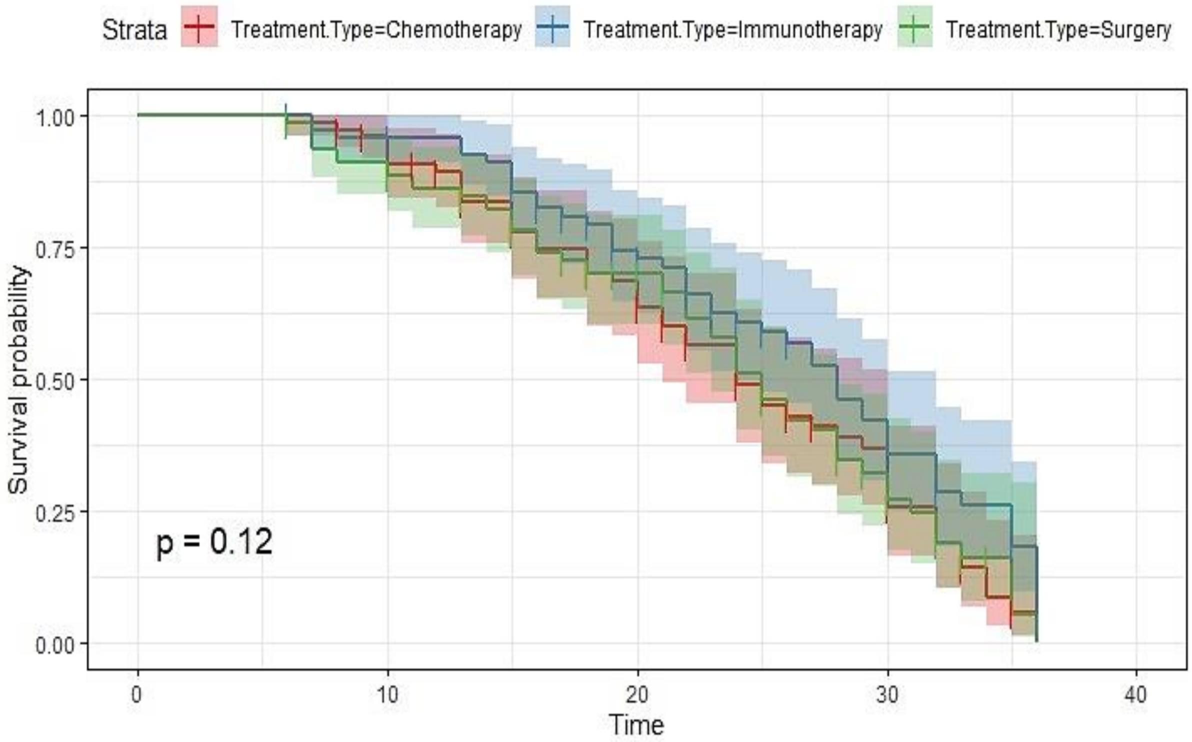

Figure 5. Kaplan-Meier curve on treatment type

The Kaplan-Meier curve from Figure 5 shows no statistically significant difference in survival outcomes between chemotherapy, immunotherapy, and surgery, with a p-value of 0.12.

Table 7. Summary of log-rank test covariates.

Covariates	Test Statistic	P-value
Smoking status of breast cancer patients	0.3	0.58
Occupation of breast cancer patients	19.9	0.001
Stages of breast cancer patients	2.2	0.5
Treatment type of breast cancer patients	4.2	0.12

Table 7 summarizes the covariates and the log-rank test results indicate that smoking status (p = 0.58) and stages of breast cancer (p = 0.50) show no significant differences in survival probabilities, while occupation (p = 0.001) demonstrates a significant difference, and treatment type (p = 0.12) shows no significant differences.

3. 3. Semi parametric model

Cox proportional hazard model

Table 8. Results for Cox proportional hazard model.

Covariates	Hazard Ratio	Std.Error	Z	P Value	[Conf. Interval]
Age	1.0038	0.0050	0.759	0.4477	[0.9941 1.0136]
Sex (Female)	0.5694	0.3852	1.462	0.1438	[0.2676 1.2116]
Occupation 2	1.7025	0.2677	1.988	0.0468	[1.0075 2.8769]
Occupation 3	1.3863	0.4916	0.664	0.5065	[0.5289 3.6336]
Occupation 4	1.1745	0.4330	0.371	0.7104	[0.5026 2.7447]
Occupation 5	2.3307	0.7290	1.161	0.0796	[0.9315 3.5673]
Occupation 6	1.8229	0.3425	1.753	0.0796	[0.4235 1.3077]
Breast Cancer Stages I	0.8775	0.2669	-0.490	0.6244	[0.5201 1.4805]
Breast Cancer Stages II	0.6375	0.2702	-1.666	0.0957	[0.3754 1.0826]
Breast Cancer Stages III	1.0169	0.2575	0.065	0.9481	[0.6142 1.6836]
Smoking. Status (Non-Smoker)	0.5836	0.1743	-3.091	0.0020	[0.4147 0.8212]
Treatment Type (Chemotherapy)	0.8062	0.2100	-1.026	0.3050	[0.5342 1.2167]
Treatment Type (Immunotherapy)	0.6088	0.2091	-2.373	0.0176	[0.4041 0.9172]

Table 8 reveals the Cox proportional hazard model results show the effect of different covariates on hazard ratios. Age has no significant impact (HR: 1.003, p = 0.448) while being female reduces the hazard but is not significant (HR: 0.569, p = 0.144). Occupation 2 significantly increases the hazard (HR: 1.703, p = 0.047).

Breast cancer stage and smoking status significantly affect the outcome, with non-smokers having a reduced hazard (HR: 0.584, p = 0.002). Immunotherapy significantly reduces the hazard (HR: 0.609, p = 0.018). Other factors like chemotherapy, sex, and most occupations were not significant.

3. 4. Parametric model

Table 9. Exponential distribution.

Covariates	Hazard Ratio	Value	Std. Error	Z	P Value	[Conf. Interval]
Intercept	21.8217	3.0829	0.5172	5.96	2.5e-09	[2.0693 4.0965]
Age	0.9995	-0.0005	0.0050	-0.11	0.913	[-0.0102 0.0091]
Sex (Female)	1.3579	0.3060	0.3772	0.81	0.417	[-0.4335 1.0454]
Occupation 2	0.8093	-0.2115	0.2632	-0.80	0.422	[-0.7274 0.3044]
Occupation 3	0.8969	-0.1087	0.4762	-0.23	0.819	[-1.0422 0.8248]
Occupation 4	0.9481	-0.0533	0.4303	-0.12	0.901	[-0.8967 0.7900]
Occupation 5	0.6364	-0.4519	0.7241	-0.62	0.533	[-1.8710 0.9673]
Occupation 6	0.9232	-0.0799	0.3354	-0.24	0.812	[-0.7374 0.5775]
Breast Cancer Stage I	0.9704	-0.0300	0.2648	-0.11	0.910	[-0.5491 0.4890]
Breast Cancer Stage II	1.0571	0.0555	0.2577	0.22	0.829	[-0.4496 0.5607]
Breast Cancer Stage III	0.8271	-0.1898	0.2508	-0.76	0.449	[-0.6814 0.3018]
Smoking. Status (Non-Smoker)	1.3455	0.2967	0.1679	1.77	0.077	[-0.0325 0.6260]
Treatment Type (Chemotherapy)	0.9819	-0.0182	0.2010	-0.09	0.928	[-0.4123 0.3757]
Treatment Type (Immunotherapy)	1.1571	0.1459	0.2022	0.72	0.471	[-0.2504 0.5423]

Table 9 reveals results from the exponential distribution model used in evaluating the relationship between various covariates and the hazard ratio for breast cancer patients. A hazard ratio greater than 1 implies increased risk, while values less than 1 suggest reduced risk. The intercept, with a hazard ratio of 21.82 ($p < 0.001$), is significant, indicating a high baseline hazard. Covariates like age (HR = 0.9995, p = 0.913), Sex (HR = 1.36, p = 0.417), and different Occupation categories show non-significant associations with the hazard of the event. Smoking status approaches significance (p = 0.077), suggesting a possible relationship. Other factors like treatment types and breast cancer stages show no strong evidence of association.

Table 10. Log-Normal distribution model

Covariates	Hazard Ratio	Value	Std.Error	Z	P Value	[Conf. Interval]
Intercept	18.9835	2.9435	0.2358	12.48	<2e-16	[2.0693 4.0965]
Age	0.9976	-0.0024	0.0022	-1.11	0.2672	[-0.0102 0.0091]
Sex (Female)	1.10293	0.0980	0.17291	0.57	0.5710	[-0.4335 1.0454]
Occupation 2	0.9050	-0.0998	0.1186	-0.84	0.3998	[-0.7274 0.3044]
Occupation 3	0.8188	-0.2000	0.2035	-0.98	0.3264	[-1.0422 0.8248]
Occupation 4	0.8952	-0.1107	0.1829	-0.61	0.5449	[-0.8967 0.7900]
Occupation 5	0.9015	-0.1037	0.3562	-0.29	0.7709	[-1.8710 0.9673]
Occupation 6	0.8426	-0.1713	0.1414	-1.21	0.2256	[-0.7374 0.5775]
Breast Cancer Stage I	1.0752	0.0726	0.0726	0.65	0.5184	[-0.5491 0.4890]
Breast Cancer Stage II	1.2577	0.2293	0.1083	2.12	0.0343	[-0.4496 0.5607]
Breast Cancer Stage III	1.0110	0.01092	0.1071	0.10	0.9188	[-0.6814 0.3018]
Smoking. Status (Non-Smoker)	1.2286	0.2058	0.0737	2.79	0.0053	[-0.0325 0.6260]
Treatment Type (Chemotherapy)	1.0245	0.0242	0.0883	0.27	0.78430	[-0.4123 0.3757]
Treatment Type (Immunotherapy)	1.1827	0.1678	0.1678	0.0892	0.0598	[-0.2504 0.5423]

Table 10 Presents the results of the Log-Normal distribution model assessing the relationship between covariates and the hazard ratio for breast cancer patients. The intercept is significant (HR = 18.98, p < 0.001), indicating a high baseline hazard. Age (HR = 0.998, p = 0.2672) and Sex (HR = 1.10, p = 0.5710) show non-significant associations.

Breast cancer stages I and III also show non-significant results, but Stage II is significant (HR = 1.26, p = 0.0343), indicating higher risk. Smoking status (HR = 1.23, p = 0.0053) shows a significantly increased hazard, while treatment types are not statistically significant. This suggests that smoking and Stage II breast cancer are notable risk factors under the Log-Normal model.

3. 5. Evaluating the models used in the study of survival analysis using model comparison

Table 11. Model evaluation

MODEL	AIC	BIC
Cox ph model	1385.218	1424.698
Exponential model	1402.989	1464.875
Log-normal model	1255.282	1302.461

Table 11 compares survival models using AIC and BIC values, showing that the Log-Normal model performs best, with the lowest AIC (1255.282) and BIC (1302.461), indicating a better fit while accounting for model complexity. The Cox Proportional Hazards model ranks second with an AIC of 1385.218 and a BIC of 1424.698. The Exponential model, with the highest AIC (1402.989) and BIC (1464.875), fits the data least effectively. Overall, the Log-Normal model provides the best balance between accuracy and simplicity in this analysis.

4. CONCLUSIONS

The analysis of breast cancer patient data reveals significant insights into survival probabilities influenced by various factors. The majority of respondents were female (96%), with a notable distribution across cancer stages: Stage II had the highest representation at 41%, while Stage IV had the lowest at 13%.

The Kaplan-Meier curve indicated a gradual decline in survival probability over 35 months, with a marked drop to around 50% by 18-20 months, ultimately nearing zero by the study's end. Notably, the log-rank tests highlighted that smoking status and occupation significantly impacted survival outcomes, with p-values of 0.006 and 0.001 respectively. In contrast, factors such as Breast Cancer Stage and treatment type did not show significant effects on survival probabilities.

The findings suggest that lifestyle factors like smoking and occupational exposure may play critical roles in breast cancer survival, while traditional clinical factors like stage and treatment type may not be as influential as previously thought, the Log-Normal model performs best, with the lowest AIC (1255.282) and BIC (1302.461), indicating a better fit while accounting for model complexity.

There is a need for further investigation into how these lifestyle factors can be mitigated or managed to improve patient outcomes. It is recommended that healthcare providers focus on educating patients about the risks associated with smoking and consider occupational health assessments as part of comprehensive cancer care.

Future research should explore targeted interventions for high-risk occupational groups and develop strategies to support smoking cessation among breast cancer patients to enhance survival rates.

References

[1] Barker, L. E., & O'Connell, J. (2021). The Importance of Survival Model Selection in Cancer Research: A Focus on Breast Cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 30(5), 889-895. https://doi.org/10.1158/1055-9965.EPL-20-0968

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394-424. https://doi.org/10.3322/caac.21492

[3] Carroll K.J (2021). On the use and utility of the Weibull model in the analysis of survival data. *Control Clinical Trials* 2003; 24: 682-701

[4] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-220

[5] Ghosh, A., Roy, S., & Bhattacharyya, S. (2020). Survival Analysis: A Comparative Study of Cox Proportional Hazards and Accelerated Failure Time Models. *JAMA Oncol*. 2020; 6(12): 1983-1984. doi:10.1001/jamaoncol.2020.5158

[6] Giordano, S. H., Elias, A. D., & Gradishar, W. J. (2018). NCCN Guidelines updates Breast cancer. *JNCCN–Journal of the National Comprehensive Cancer Network*, 16(5S), 605-610. https://doi.org/10.6004/jnccn.2018.0035

[7] Gomez, J. R., & Lammers, A. E. (2022). Statistical modeling of quality of life and patient-reported outcomes in breast cancer: Current practices and future directions. *Journal of Clinical Oncology*, 40(10), 1120-1129. https://doi.org/10.1200/JCO.2021.40.10_suppl.1120

[8] Kim, H. J., & Kim, H. (2021). Comparison of survival analysis methods for breast cancer: A comprehensive review. *Journal of Biomedical Informatics*, 113, 103661 .https://doi.org/10.1016/j.jbi.2020.103661

[9] Kleinbauim D.G & Klein M. Survival Analysis: A Self-Learning Text. Springer Science+Business Media, LLC, part of Springer Nature 2012. https://doi.org/10.1007/978-1-4419-6646-9

[10] Lee, E. T., & Wang, J. W. (2020). Statistical Methods for Survival Data Analysis. John Wiley & Sons, Inc. Online ISBN:9780471458548. DOI: 10.1002/0471458546

[11] Medhat Mohamed Ahmed Abdelaal & Sally Hossam Eldin Ahmed Zakria (2015). Modeling Survival Data by Using Cox Regression Model. *American Journal of Theoretical and Applied Statistics* 4(6): 504-512

[12] Miller, A. B., & Hodge, W. (2020). Evaluating the Assumptions of Survival Analysis: A Practical Guide for Researchers. *Journal of Clinical Epidemiology*, 123, 51-57. https://doi.org/10.1016/j.jclinepi.2019.12.010

[13] Miller, T. R., Olsen, K., & Williams, P. (2023). Parametric models in oncology: A review with applications to breast cancer survival data. *Statistics in Medicine*, 42(1), 52-68. https://doi.org/10.1002/sim.9323

[14] Nassif EF, Cope B, Traweek R, Witt RG, Erstad DJ, Scally CP, Thirasastr P, Zarzour MA, Ludwig J, Benjamin R, Bishop AJ, Guadagnolo BA, Ingram D, Wani K, Wang WL, Lazar AJ, Torres KE, Hunt KK, Feig BW, Roland CL, Somaiah N, Keung EZ. Real-world use of palbociclib monotherapy in retroperitoneal liposarcomas at a large volume sarcoma center. *Int J Cancer.* 2022 Jun 15; 150(12): 2012-2024. doi: 10.1002/ijc.33956

[15] Nguyen, T., & Chen, R. (2021). Understanding censoring mechanisms and their impact on survival analysis models in clinical trials. *Statistical Methods in Medical Research*, 30(4), 905-916. https://doi.org/10.1177/0962280220984105

[16] Ryosuke Fujii (2023). Visualization of Relative Measures of Association: Points and Error Bars With an Appropriate Axis Scale. *Journal of Epidemiology* 33(9). https://doi.org/10.2188/jea.JE20230052

[17] Rahman, H., Liu, Q., & Zhang, Y. (2022). A comparative study of survival models for breast cancer patients: Cox proportional hazards, parametric models, and their performance. *Journal of Cancer Research and Clinical Oncology*, 148(5), 1012-1024. https://doi.org/10.1007/s00432-021-03650-9

[18] Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. *Anesth Analg.* 2018 Sep; 127(3): 792-798. doi: 10.1213/ANE.0000000000003653

[19] Smith, R. E., Anderson, R. E., & Smith, M. L. (2021). Survival Analysis in Medicine and Epidemiology: A Comprehensive Overview. *Journal of Epidemiology & Community Health*, 75(5), 445-452. doi:10.1136/jech-2020-213209

[20] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. DOI: 10.3322/caac.21660

[21] Wang, K., & Wei, Z. (2020). The Cox Proportional Hazards Model: Applications and Practical Issues. *Current Problems in Cancer*, 44(5), 100501. https://doi.org/10.1016/j.currproblcancer.2020.100501

[22] Zhu, Y., Li, F., & Chen, X. (2021). Assessing the performance of survival models in breast cancer research: Cox PH, exponential, Weibull, and log-normal comparisons. *BMC Medical Research Methodology*, 21(48), 1-14. https://doi.org/10.1186/s12874-021-01244-8