Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER
PUBLISHER

Research Article

Statistical Evaluation of Survival Rates in Lung Cancer Utilizing Gaussian and Logistic Regression Techniques

Pitta Shankaraiah, Mokesh Rayalu. G*

School of Advanced sciences, Vellore Institute of Technology, Vellore, India
E-mail: mokesh.g@vit.ac.in

Received: 12 August 2024; **Revised:** 25 September 2024; **Accepted:** 22 October 2024

Abstract: Cancer is the second most prevalent cause of mortality globally as per the World Health Organization. Among the various types of cancer, lung cancer is particularly fatal and ranks third in terms of frequency. Its impact on healthcare systems and individuals' quality of life is enormous. This sort of cancer is particularly problematic because it has a very poor survival rate compared to other types of cancer. The focus of the present research is to examine the correlation between lung cancer and survival time, in addition to the different characteristics of the cancer dataset. The purpose of the present investigation is to determine the optimal modeling strategy for accurately assessing the survival probabilities and other statistical measures. A set of Gaussian and Logistic parametric regression survival models to calculate probability values, average survival time, and other relevant statistical metrics have been used in the present research study. The data of 168 patients and nine essential variables related to advanced lung cancer, including age, gender, and other clinical factors have been included in the study. The proposed estimation methods are compared by assessing significant factors, such as mean survival probability, mean cumulative survival probability, and model fit indices viz, the Akaike Information Criterion and Bayesian Information Criterion. The family of Logistic Regression models exhibited higher performance across these parameters, reflecting their resilience and appropriateness for this particular set of survival data.

***Keywords*:** Correlation, Cox Proportional Hazard Model, Model performance tests, Logistic Regression, Gaussian Regression, Statistical metrics

MSC: 60E05, 62B15, 62F30, 62G05, 62P10

Abbreviation

LR, GR	Logistic Regression and Gaussian Regression
LL, LG	Log-logistic and Log-gaussian
Pdf, CDF	Probability density function and Cumulative Distribution Function
SF, HF	Survival and Hazard Function
CSF, CHF	Cumulative Survival Function and Cumulative Hazard Function
MSP	Mean Survival Probability
MCSP	Mean Cumulative Survival Probability

DOI: https://doi.org/10.37256/cm.5420245479

AIC, BIC	Akaike and Bayesian Information Criteria
Cox PH	Cox Proportional Hazard

1. Introduction

Cancer is one among of the most frequent causes of non-accidental deaths. Multiple research studies have shown that lung cancer is the leading cause of cancer-related deaths in the global scenario which may lead to a medical conclusion for lung cancer [1]. Usually, at the final phase of the disease, the patient will be aware of it. The primary cause of lung cancer is definitely smoking which is harmful due to it prevalence and final screening stage, and also limits strategies for treatment and lowers survival chances [2]. Lung cancer is a highly prevalent type of cancer across the globe. Since majority lung cancer victims are discovered at the middle stage or final stage and have bad predictions, as the disease's early signs are not always evident . One successful strategy to lower mortality in high-risk patients is lung cancer screening [3]. Lung cancer fatalities across all subtypes between 2022 and 2023 in which there were 130,180 fatalities from 236,740 cases of lung cancer in 2022. Because so many people die from cancer each year, lung cancer is still a major concern for the world. It is projected in the United States alone that, 238,340 persons might receive a lung cancer diagnosis by 2023 [4]. Preliminarily the data shows that 1 in 16 people will receive a lung cancer diagnosis at a moment in their lives, with men being more likely to be diagnosed (1 in 16) than women (1 in 17). Smoking is the main cause of lung cancer, responsible for over 80% of deaths which is significant, therefore, the rest 20% of lung cancer deaths involve non-smokers [5].

The primary goals of this study are:

(i) to identify the pairwise relationship by considering all the variables using pearson correlation.

(ii) to ensure the significant components using p-values, Cox Proportional Hazard (Cox PH) model has been used.

(iii) to obtain the mean survival time, Mean Survival Probability (MSP), and Mean Cumulative Survival Probability (MCSP) by using parametric regression survival models on the data of lung cancer patients and

(iv) finally, to find the suitable and best-performing the Logistic Regression (LR) and Gaussian Regression (GR) survival methods through Akaike Information Criterion (AIC), and Bayesian Information Criteria (BIC) values.

The paper has been structured as follows: Section 2 provides a review of the studies written by the many authors that are pertinent to the investigation of lung cancer. The details of the data collection, employed techniques and quantitative investigation of the variables in lung cancer are covered in Section 3. Section 4 discusses an existence and applicability of the mathematical framework. In Section 5 and 6 presents the facts gathered, the discussion and conclusion.

2. Literature survey

A study examined by Tirzite et al. [6] LR strategy has been used to separate cancer victims from regular people by examining patient exhaled breath samples. An electronic nasal device and LR analysis are used in the study to examine exhaled breath samples from 223 non-cancer patients and 252 cancer patients. Findings indicate that the sensitivity was 95.8% for smokers and 96.2% for non-smokers, while the specificity was 90.6% for non-smokers and 92.3% for smokers. The effects help to distinguish lung cancer patients from healthy people. To examine the prevalence of lung cancer and to indicate important causes, a study has been conducted by Qun and Abd Rahman [7], Which involves statistical analysis, and makes use of LR analysis model for 309 records with 16 attributes. Since the results demonstrate substantial connections with the development of lung cancer, the optimum multivariate logistic model was created by taking into account the factors with p-values less than 0.05. This study supports oncology initiatives aimed at early recognition and prevention. The study introduced by Muse et al. [8] experimented with a novel generalized Log-logistic (LL) distribution that expands upon current distributions such as Exponential, Weibull, LL, and Burr XII. using Monte Carlo simulations, real-life data evaluation, basic characteristics testing, and parameter estimates, mathematical and statistical characteristics have been demonstrated. In summary, the generalized LL model is found to be the most prosperous and appropriate model for understanding lifetime phenomena. The performance of the LL model with covariate, right, and interval-censored data using various imputation to analyze techniques, including maximum likelihood estimation and modified Cox-Snell has been explored by Lai and Arasan [9]. The dataset comprises 731 patients out of which 277 females and 454 males. It does not contain any right-censored data however it does contain 136 interval-censored observations. Modified Cox-Snell residuals performed better than conventional residuals in terms of model adequacy. Male patients had somewhat greater survival probability than female patients in the LL model, which fits the diabetes data quite well. A study of Yang et al. [3] makes use of LR analysis and Receiver Operating Characteristic curves to assess the clinical usefulness of biomarker mixtures for lung cancer screening in which 633 victims of lung cancer and 650 individuals with pneumonia involved. There were eight pairings with noticeably higher areas under the curves. The most effective screening indicators for patients with pneumonia revealed variations in sex but not age. For serological tumor screening, a two-marker combination was found to be more appropriate than a multimarker combination. In a particular study of lung cancer by Escudero-Vilaplana et al. [10] individuals in Spain with the initial stages of non-small cell lung cancer were assessed for the cost-effectiveness of atezolizumab as an adjuvant treatment. It contrasts platinum-based chemotherapy and optimal supportive treatment following resection with atezolizumab. To evaluate long-term medical and financial outcomes, a Markov model with five health stages and monthly cycles is used in the analysis. The findings demonstrate that atezolizumab, albeit at a higher cost, resulted in more life years and quality-adjusted life years when compared to optimal supportive care. The Sory Traore et al. [11] research closes the loopholes in economic analyses brought-out by progression-free survival based primary objectives in the field of cancer. To predict overall survival time and hazard rate have been combined into a univariate model with a joint Bayesian technique. The study showed that when compared to the frequentist approach, the univariate and joint Bayesian models decreased the risk of overall survival time prediction. A study of Wong et al. [12] focused on the mortality risk that varies with time after a lobectomy for lung cancer. It analyzed 2,284 patients between 2015 and 2022 using parametric survival models and a Hazard Function (HF) displayed across time in respect to the best fit statistical distribution. The findings showed that at 30, 90, and 180 days, the cumulative death rate was 1.3%, 2.9%, and 4.9% respectively. The moment's risk rate has raised during the postoperative period. It decreased sharply in the first thirty days and then become consistent at 180 days. Research conducted by Ahmed [13], employs parametric methods, including Weibull, Gumbel, Exponential, and LL, to determine the Survival Function (SF) of lung cancer patients. The Gumbel distribution model was shown to be the most successful method after evaluating the approaches with corrected, Akaike, and Bayesian informations. SF and Failure periods are inversely related; as the latter rises, the former falls.

In the LL model, the investigation mainly focuses on one covariate. However, residual analysis, external validation, and multiple covariates might not be covered in detail in this study [9]. Another study failed to address any impact of explanatory factors on the shapes of hazard rates and survival outcomes, and instead recommended extending the generalized LL distribution to account for accelerated failure times and competing risks [8]. Furthermore, a study attempts to close current gaps by examining and comparing the time-varying risk of death among various lung resection procedures [12]. The aforementioned studies use various survival models to design the Receiver Operating Characteristic and areas under the curves while examining progression-free survival, overall survival, hazard rate, risk variables, parameters, lung cancer prevalence, and the survival mean. In order to analyze the data, the study worked with LR model and GR based on significant factors, survival probability, mean cumulative probability, AIC, and BIC values.

3. Quantitative evaluation of key variables in lung cancer data

The present research study makes use of Terry Therneau's survival statistics for patients with advanced lung cancer in the North Central Cancer Treatment Group [14]. Performance scores evaluate the patient's ability to carry out regular everyday tasks. The present dataset contains 168 patients ranging from 39 to 82 years of age where survival times range from 5 days to 1,022 days at most. The 168 records of 9 columns of the data sample are included with these variables. Time is a conditional variable and all other variables are influenced, as demonstrated by the use of every unit in this data set which is enunciated in Table 1.

Table 1. The list of variables representing the lung cancer dataset

S NO.	Variables	Description
1	Time	Survival time in days
2	Status	Censoring status, (Censored 1 and Dead 2)
3	Age	Survival time in days
4	Sex	Male 1 and Female 2
5	Ph.E	Ecog performance rating; (Good 0 to Dead 5)
6	Ph.K	Physician-rated karnofsky performance rating; (Bad 0 to Good 100)
7	Pa.K	Patient-rated karnofsky performance rating; (Bad 0 to Good 100)
8	ML	Calories taken during meals
9	WL	Loss of weight throughout the past six months

The complete data analysis flow and functionality is shown in Figure 1. Data processing is necessary to sort the dataset so that it can be used for a variety of survival models viz., Cox PH model, and combinations of Logistic and Gaussian family regression models [15]. R computer programming is used to carry out the statistical analysis. Aspects including age, sex, performance ratings, caloric consumption, weight reduction, and more are included in the data collection. The effect of every variable on the patient lifespan would be better understandable with the aid of this analysis. The present investigation is on some individuals of 39 to 82 years of age, with bifurcation into male and female categories. Age and sex are important demographic characteristics that may affect survival time. Based on these criteria, variations in survival outcomes are frequently observed; younger patients and female patients tend to perform more consistently and cure faster, and the survival period for some cancer kinds is getting longer. When rating a patient's ability to execute daily tasks on a scale of 0 to 100, the performance scores (Ph.E, Ph.K) are essential. An improved score (100) denotes a better functional state, which may be associated with greater survival rates. Two nutritional status indicators that may have an impact on general health and possibly the prognosis of cancer patients are calorie intake and weight reduction [16]. The effectiveness of the Logistic and Gaussian family regression models is assessed using a patient population with advanced lung cancer to determine the statistical metrics, AIC, and BIC values that are obtained.

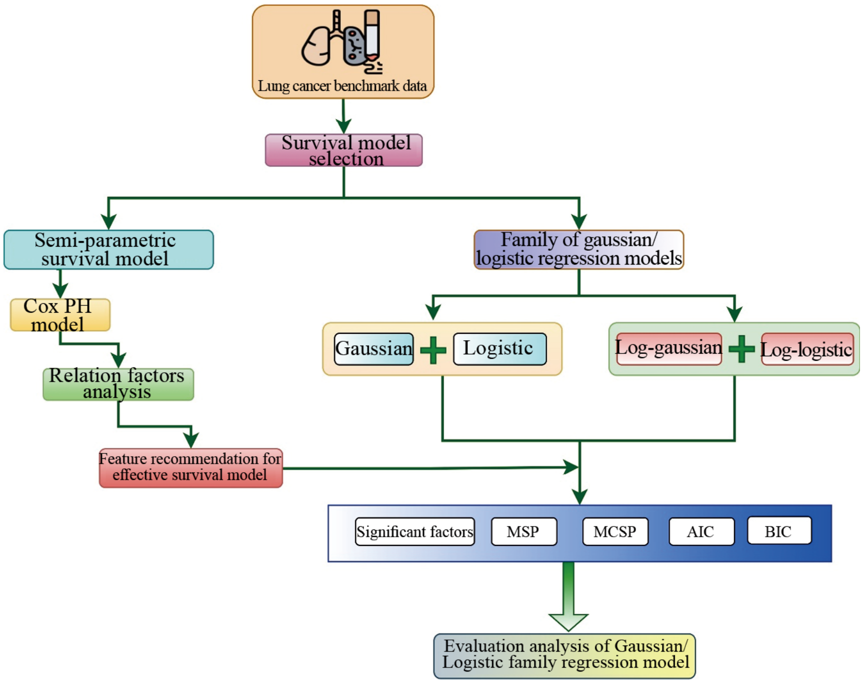
Lung cancer benchmark data
Survival model selection
Semi-parametric survival model
Family of gaussian/ logistic regression models
Cox PH model
Gaussian
Logistic
Log-gaussian
Log-logistic
Relation factors analysis
Feature recommendation for effective survival model
Significant factors
MSP
MCSP
AIC
BIC
Evaluation analysis of Gaussian/ Logistic family regression model

Figure 1. Design and structure of the proposed methodological model

The strength of the linear relationship between both variables is referred to as correlation in general. The direction of a linear relationship between both variables is measured to pairwise by using pearson correlation strategy [17]. The Pearson correlation coefficient for each pair of variables x_i and x_j is given by

$$r_{(x_i,\, x_j)} = \frac{\operatorname{cov}(x_i,\, x_j)}{\sigma_{x_i} \cdot \sigma_{x_j}};\ i,\ j = 1,\ 2,\ \ldots,\ n;\ -1 \leq r \leq 1$$

Where

$$\operatorname{cov}(x_i,\, x_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

$$\sigma_{x_i} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_i)^2}$$

$$\sigma_{x_j} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_j)^2}$$

Table 2. Pairwise analysis of correlated features in lung cancer

Pairwise correlation	Time	Status	Age	Sex	Ph.E	Ph.K	Pa.K	ML	WL
Time	1	-0.16	-0.08	0.11	-0.19	0.09	0.18	0.07	0.03
Status		1	0.16	-0.22	0.24	-0.16	-0.19	0.02	0.05
Age			1	-0.13	0.31	-0.33	-0.24	-0.24	0.05
Sex				1	-0.01	-0.02	0.07	-0.17	-0.17
Ph.E					1	-0.82	-0.54	-0.11	0.18
Ph.K						1	0.53	0.06	-0.13
Pa.K							1	0.17	-0.18
ML								1	-0.11
WL									1

The results indicated in Table 2 has been obtained by using the pearson formula as shown above. And it represents a slight decrease in the probability of the event with time in case of age, Ph.E, and Pa.K. On the other hand, there is a low association with time in the case of Ph.K, ML, WL, and individuals' gender, as indicated by a low positive correlation.

4. Applied statistical models and inference techniques

The hazard rate in a semi-parametric model has been considered with a finite-dimensional parameter of interest which may be expressed as a baseline hazard multiplied by a term that only interacts towards the covariates [18, 19]. When estimating the length of time to a particular instance based on the values of provided covariates, the Cox regression model works effectively [20].

Here is the formula for the Cox PH:

$$h(t \mid X_i) = h_0(t) \cdot \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_7 X_{i7})$$

$$h(t \mid X_i) = h_0(t) \cdot \exp(X_i \cdot \beta) \tag{1}$$

where $h(t \mid X_i)$ represent the HF at time t, and β_i are the parameters with the predictor variables X_i. In equation [1], the baseline hazard constant $h_0(t)$ is identical. The only difference is subjects HF comes from the baseline scaling factor $\exp(X_i.\beta)$. In majority of instances, there is a correlation between the covariates as shown in Table 2, which requires an appropriate selection of variables procedure in the present situation.

Parametric models require a completely stated HF and provide greater statistical testing capabilities compared to semi-parametric or non-parametric models, provided that a suitable model can be identified and the assumptions of parametric modeling are satisfied [21]. Presently, it has become essential to find studies on consistency in addition to employing survival analysis to analyze clinical and epidemiological data [22]. Researchers are continuously working with different kinds of "parametric distributions" when analyzing survival data.

Meanwhile, parametric techniques can only be employed in this study for comparison. The LR is a predictive analysis. A statistical method called LR is applied in predictive modeling in order to explain data and clarify the interaction between the independent and dependent variables [7].

Multiple logistic regression is an extension of LR that allows more than one predictor variable. The model predicts the probability of a binary outcome (censored/death) based on multiple independent variables.

Let Y be the outcome variable, and let $X_1, X_2, \ldots, X_7$ be the predictor variables. The multiple logistic regression is given by:

$$logit(P(Y|X_1, X_2, \ldots, X_7)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7$$

where:

β_0 is the intercept term.

$\beta_1, \beta_2, \ldots, \beta_7$ indicates coefficient values for predictor variables $X_1, X_2, \ldots, X_7$.

The Logistic function transforms the log back to a probability:

$$P(Y|X_1, X_2, \ldots, X_7) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_7 X_7))}$$

To estimate parameters by using maximize this likelihood function.

Let $(y_i, x_{i1}, x_{i2}, \ldots, x_{i7})$ be the i-th observation in the dataset, where y_i is the binary outcome and $x_{i1}, x_{i2}, \ldots, x_{i7}$ are the predictor values for that observation. The likelihood function for n observations is:

$$L(\beta_0, \beta_1, \ldots, \beta_7) = \prod_{i=1}^{7} P(Y = y_i|X_{i1}, X_{i2}, \ldots, X_{i7})^{y_i} \cdot (1 - P(Y = y_i|X_{i1}, X_{i2}, \ldots, X_{i7}))^{1-y_i}$$

Substitute the Logistic function into the likelihood function:

$$L(\beta_0,\ \beta_1,\ \ldots,\ \beta_7)=\prod_{i=1}^{7}\left(\frac{1}{1+\exp(-(\beta_0+\beta_1x_{i1}+\cdots+\beta_7x_{i7}))}\right)^{y_i}\cdot\left(1-\frac{1}{1+\exp(-(\beta_0+\beta_1x_{i1}+\cdots+\beta_7x_{i7}))}\right)^{1-y_i}$$

Taking the natural logarthim of likelihood function, after simplifying:

$$\ell(\beta_0,\ \beta_1,\ \ldots,\ \beta_7)=\sum_{i=1}^{7}[y_i(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_7x_{i7})-\log(1+\exp(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_7x_{i7}))]$$

The log-likelihood function is maximized by the values that correspond to the maximum likelihood estimates of parameters $\beta_0,\ \beta_1,\ \ldots,\ \beta_7$ for each parameter, this necessitates taking the partial derivatives of the log-likelihood function and converting to zero.

$$\frac{\partial\ell}{\partial\beta_j}=0 \text{ for } j=0,\ 1,\ \ldots,\ 7 \tag{2}$$

The equation (2) has been used for computing the predictors coefficients with the help of LR, GR, LL,and LG techniques.

4.1 *Exploring the logistic distribution*

A random variable time t follows a Logistic distribution with shape parameter μ and scale parameter S then Pdf of the Logistic distribution is:

$$f(t)=\frac{\exp\left(-\frac{t-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{t-\mu}{s}\right)\right)^2};\ t>0.$$

To derive the CDF $F(t)$, we integrate the Pdf from $-\infty$ to t:

$$F(t)=\frac{1}{1+\exp\left(-\frac{t-\mu}{s}\right)}$$

Using the relationship of SF and CDF

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

Using the CDF of the Logistic distribution:

$$S(t) = 1 - \frac{1}{1 + \exp\left(-\frac{t-\mu}{s}\right)}$$

$$S(t) = \frac{\exp\left(-\frac{t-\mu}{s}\right)}{1 + \exp\left(-\frac{t-\mu}{s}\right)}$$

After Simplifying the expression:

$$s(t) = \frac{1}{1+\exp\left(-\frac{t-\mu}{s}\right)} \tag{3}$$

Using the relationship of Pdf and SF then the HF is going to be

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

By substituting the Pdf and SF in above relation, after simplification the expression would be:

$$h(t) = \frac{1}{s\left(1 + \exp\left(-\frac{t-\mu}{s}\right)\right)} \tag{4}$$

After applying integral to the above Eq. then the Cumulative Hazard Function (CHF) is

$$H(t) = \int_{-\infty}^{t} \frac{1}{s\left(1 + \exp\left(-\frac{u-\mu}{s}\right)\right)} du$$

Let $v = \frac{u-\mu}{s}$, then $du = s\,dv$:

$$H(t) = \int_{-\infty}^{\frac{t-\mu}{s}} \frac{1}{1 + \exp(-v)} dv$$

The integral of $\frac{1}{1+\exp(-v)}$ is known to be $\ln(1+\exp(v))$:

$$H(t) = \ln(1 + \exp(v))|_{-\infty}^{\frac{t-\mu}{s}}$$

Since $\exp(-\infty) = 0$:

$$H(t) = \ln\left(1 + \exp\left(\frac{t-\mu}{s}\right)\right) \tag{5}$$

4.2 *Mathematical derivation of log-logistic distribution*

By applying logarthim to the corresponding independent variable 't' in the Logistic distribution, the LL distribution can be obtained. The LL distribution has two parameters as scale parameter β and shape parameter α. Then the Pdf of LL would be

$$f(t) = \frac{(t/\beta)^{\alpha-1}\alpha/\beta}{[1+(t/\beta)^{\alpha}]^2}$$

To derive the CDF $F(t)$, we integrate the Pdf from 0 to t:

$$F(t) = \frac{1}{1+\left(\frac{t}{\beta}\right)^{-\alpha}}$$

Using the relationship of SF and CDF

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

$$S(t) = 1 - \frac{1}{1+(t/\beta)^{-\alpha}}$$

After Simplifying the expression:

$$s(t) = \frac{1}{1+(t/\beta)^{\alpha}} \tag{6}$$

Using the relation of Pdf and SF then the HF is going to be

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

Substitute the Pdf and SF of the LL distribution in above relation, then after the simplification the equation form is:

$$h(t) = \frac{(t/\beta)^{\alpha-1}\alpha/\beta}{[1+(t/\beta)^{\alpha}]} \tag{7}$$

Using the relationship between the CHF and the SF, we know that:

$$H(t) = -\ln(S(t))$$

$$H(t) = -\ln(\frac{1}{1+(t/\beta)^{\alpha}})$$

$$H(t) = \ln\left(1+\left(\frac{t}{\beta}\right)^{\alpha}\right) \tag{8}$$

4.3 *Investigating the gaussian distribution*

If time 't' is independent continuous random variable with mean μ is a shape parameter and σ is the scale parameter then Pdf of the Gaussian distribution is:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right);\ \mu,\ \sigma > 0;\ -\infty < t < \infty$$

To derive the CDF from the computed Pdf

$$F(t) = 1 - \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

By using relationship of SF and CDF then the formulated equation would be $s(t) = 1 - F(t)$

$$s(t) = 1 - \left(1 - \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\right)$$

Then

$$s(t) = \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{9}$$

By using relationship between Pdf and SF then the computed HF is

$$h(t) = \left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{10}$$

Finally the CHF is

$$H(t) = 1 - \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{11}$$

4.4 *Mathematical derivation of Log-gaussian distribution*

If t =log(t) follows a Log-gaussian (LG) ditrinbution with parameters μ and σ. The Pdf of the LG distribution is:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right);\ t > 0.$$

To derive CDF $F(t)$ by using the integral of the Pdf with lower and upper limits 0, t.

$$F(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

where Φ is the CDF of the standard Gaussian distribution.

Using the relationship of SF and CDF

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

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{12}$$

By using relationship of Pdf and SF the computed HF is

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

After the simplification by substituting the Pdf and SF in the above relation the computed HF is:

$$h(t) = \frac{1}{s(t)\sigma t}\Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{13}$$

Using the relationship between the CHF and the SF, we know that:

$$H(t) = -\ln(S(t))$$

Substitute the SF of the LG distribution:

$$H(t) = -\ln\left(1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)\right)$$

$$H(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{14}$$

The Mean Survival Probability (MSP) represents the average likelihood of an individual surviving beyond a specified time t across the entire time period. The MSP can be determined using the SF $S(t)$, which denotes the likelihood of living beyond a particular time t.

The MSP over a time period T is given by:

$$\text{MSP} = \frac{1}{T}\int_0^T s(t)\,dt \tag{15}$$

The Mean Cumulative Survival Probability (MCSP) reflects the cumulative survival experience over time period $[0, T]$ is

$$\text{MCSP} = \frac{1}{T}\int_0^T \left(\int_0^t S(t)\,dt\right) dt \tag{16}$$

where $S(t)$ is the SF.

AIC and BIC quantitatively balance model fit and complexity, making them useful tools for distribution selection [23]. By offering the lowest values when comparing several distributions, these criteria help choose the optimal distribution. The formulas for AIC and BIC are :

$$\text{AIC} = -2\ln(L) + 2k \tag{17}$$

$$\text{BIC} = -2\ln(L) + k\ln(n) \tag{18}$$

Where:

L is the maximum likelihood of the model.

k is the number of parameters in the model.

n is the sample size.

5. Results and discussion

In the comparative analysis, several formulae have been framed by using equation (2), and utilized to ascertain the coefficient values associated with the endogenous variables as listed under Tables 3 and 4. These coefficients are significant markers of the relationships among these variables and their impact on the prognosis of individuals with advanced lung cancer.

Table 3. Analysis of Gaussian and Logistic coefficients for lung cancer data

Features	Gaussian model			Logistic model		
	Coefficient value	Standard error	Odds ratio	Coefficient value	Standard error	Odds ratio
Intercept	535.16 ($p = 0.1153$)	339.8684	$2.6113e^{232}$	149.6319 ($p = 0.6677$)	348.5748	$9.6454e^{64}$
age	-0.822 ($p = 0.7197$)	2.2904	$4.3953e^{-1}$	0.4758 ($p = 0.8295$)	2.2091	$1.6092e^{0}$
sex	111.52 ($p = 0.0069$)	41.2817	$2.6946e^{48}$	129.6679 ($p = 0.0016$)	41.0732	$2.0609e^{56}$
ph.ecog	-137.86 ($p = 0.0029$)	46.3138	$1.3426e^{-60}$	-107.6973 ($p = 0.0233$)	47.4595	$1.6891e^{-47}$
ph.karno	-4.52 ($p = 0.0860$)	2.6334	$1.0876e^{-2}$	-1.6580 ($p = 0.5619$)	2.8581	$1.9052e^{-1}$
pat.karno	2.40 ($p = 0.1136$)	1.5176	$1.1034e^{1}$	2.0651 ($p = 0.1571$)	1.4594	$7.8859e^{0}$
meal.cal	0.0361 ($p = 0.4656$)	0.0494	$1.0367e^{0}$	0.0598 ($p = 0.2116$)	0.0478	$1.0616e^{0}$
wt.loss	1.6164 ($p = 0.2551$)	1.4204	$5.0348e^{0}$	1.6717 ($p = 0.2309$)	1.3954	$5.3212e^{0}$

Table 4. Analysis of LG and LL coefficients for lung cancer data

Features	Log-gaussian model			Log-logistic model		
	Coefficient value	Standard error	Odds ratio	Coefficient value	Standard error	Odds ratio
Intercept	6.271687 $p = 3.1e^{-05}$	1.5052	529.3697	5.113689 ($p = 0.00023$)	1.3878	166.2826
age	-0.013269 ($p = 0.1902$)	0.0101	0.9868	-0.002237 ($p = 0.80077$)	0.0088	0.9977
sex	0.478505 ($p = 0.0085$)	0.1819	1.6136	0.483517 ($p = 0.00271$)	0.1612	1.6217
ph.ecog	-0.465406 ($p = 0.0232$)	0.2050	0.6278	-0.416993 ($p = 0.02779$)	0.1895	0.6590
ph.karno	-0.011976 ($p = 0.3063$)	0.0117	0.9880	-0.004924 ($p = 0.66312$)	0.0113	0.9950
pat.karno	0.008712 ($p = 0.1936$)	0.0067	1.0087	0.008223 ($p = 0.17092$)	0.0060	1.0082
meal.cal	0.000279 ($p = 0.2001$)	0.0002	1.0002	0.00020 ($p = 0.30232$)	0.0001	1.0001
wt.loss	0.008169 ($p = 0.1949$)	0.0063	1.0082	0.007245 ($p = 0.18601$)	0.0054	1.0072

Coefficient value represents the relationship between the predictor and the outcome. Standard error reflects the variability of the coefficient estimate. Smaller values indicate more precise estimates, while larger values indicate more uncertainty. Odds ratio describes how the odds of an event change with a one-unit increase in the predictor variable.

Table 3, upon considering both the models, sex and ph.ecog are identified as statistically significant predictors of survival based on p value. Sex has a positive coefficient indicating that gender significantly influences survival rate, with a large odds ratio suggesting a strong impact on survival rate. Similarly, the ph.ecog variable exhibits a negative coefficient in both the models indicating a worse performance status which significantly decreases survival rate. Similarly in Table 4, considering the LG and LL models, the variables sex and ph.ecog emerge as the most significant predictors of survival rate, in addition, the intercept indicates a strong baseline survival probability when all variables are zero. Other variables such as age, ph.karno, pat.karno, meal calories, and weight loss are not statistically significant in either of the model based on p (> 0.05) value. Thus, sex and performance status are the key determinants of survival rate outcomes in lung cancer patients in this analysis. The goodness of fit for Logistic and Gaussian models has been determined by using Anderson-Darling and Cramer Von-Mises tests. The obtained test's results shown in Table 5.

Table 5. Goodness of Fit Tests for Different Models

Models	Anderson-Darling		Cramer Von-Mises	
	Statistic	P-value	Statistic	P-value
Logistic	2.4534	$3.164e^{-6}$	0.45796	$6.724e^{-6}$
Log-logistic	1.1351	0.00556	0.20874	0.004105
Gaussian	2.1841	$1.449e^{-5}$	0.3939	$3.001e^{-5}$
Log-gaussian	0.0802	0.03691	0.15069	0.02303

To highlight the information from equations (3-5), (9-11) the following Figure 2 is created using R software. Similarly, using equations (6-8), (12-14) Figure 3 is created. The graphs for $S(t), h(t)$, and $H(t)$ in the LR and GR families are depicted in these illustrations.

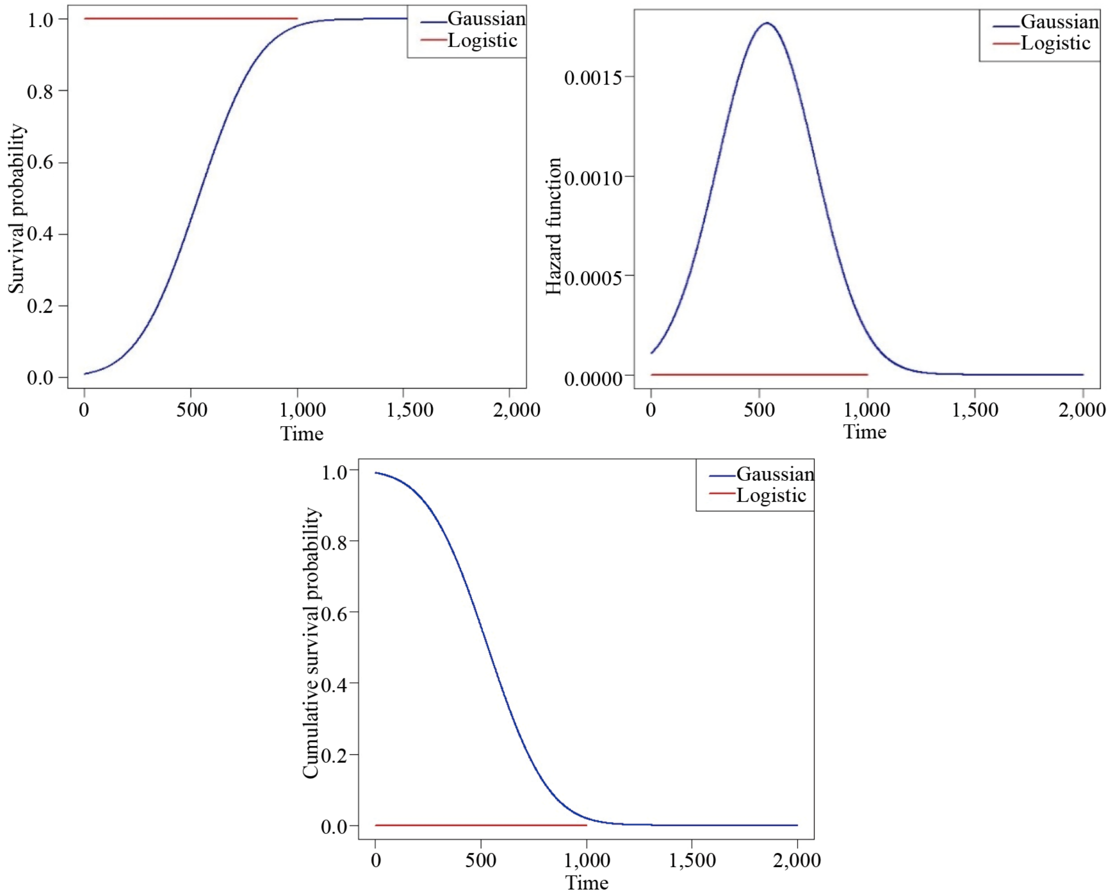

Figure 2. The sub figures represents Computed SF, HF, CSF of Gaussian and Logistic

Figure 2 shows that survival odds increasing until stabilization. The Gaussian model rises steadily, peaks at 1,500 units and then stabilizes. The Gaussian HF and CSF peak at approximately 500-time units, and subsequently decrease steeply approaching zero as time advances. The Logistic model practically constant at 1, indicating no changes in survival probability over a period of time. The HF and CSF are exhibit a fairly constant, virtually flat trajectory throughout the time, signifying a consistent risk of the event occurring at any moment.

From Figure 3, it is observed that according to the LL model, survival probability rises rapidly to towards 1 before stabilizing for the rest of the observed period. It shows that the LL model grows survival probability faster than LG model in their behavior over a period of time. Comparing the LL and LG models based on time-dependent HF, the LG model starts with a strong hazard rate at a peak stage and then rapidly declines. On the contrary, the LL model reveals a more stable and flatter hazard rate with a minor variance at early stages but a low and steady rate throughout the timeline.

The LL model predicts early risk followed by stabilization, while the LG model predicts steady risk reduction. Each model provides varied failure rates and long-term survival insights depending on the data application. The Logistic distribution exhibits a constant profile, signifying a steady survival rate throughout the time. On the other hand the Logistic distribution exhibits a practically constant HF, indicating that the risk is uniformly distributed across the time. The Logistic CSF, however, remains low, indicating a negligible survival probability from the outset.

The functions of survival models are depicted in the graphs, which include the changes in relationships among time, survival probability, and hazard rates that the model's equations require. By giving a deeper understanding of the time development of survival probabilities and hazards, such curves are helpful in the analysis and communication of the predictive model's conclusions by the researchers. With a range of 0 to 1, the probability of survival rate is an important parameter. By allowing for a comparison analysis, these values make it clear that, the prediction model provides the highest cumulative survival probabilities entirely with respect to time. As a result of the GR and LR survival models under consideration, Table 6 displays the MSP, MCSP, AIC and BIC values obtained by using the equations (15-18). The study uses the AIC and BIC values, which are quantitative measures of model effectiveness and difficulty, to evaluate and to select the most appropriate model based on statistical indicators and anticipated accuracy. The LR model is a best option as it has the fewest significant components of 2, an MSP of 1, an MCSP of 0, and the lowest values of AIC (1,692.323) and BIC (1,664.88).

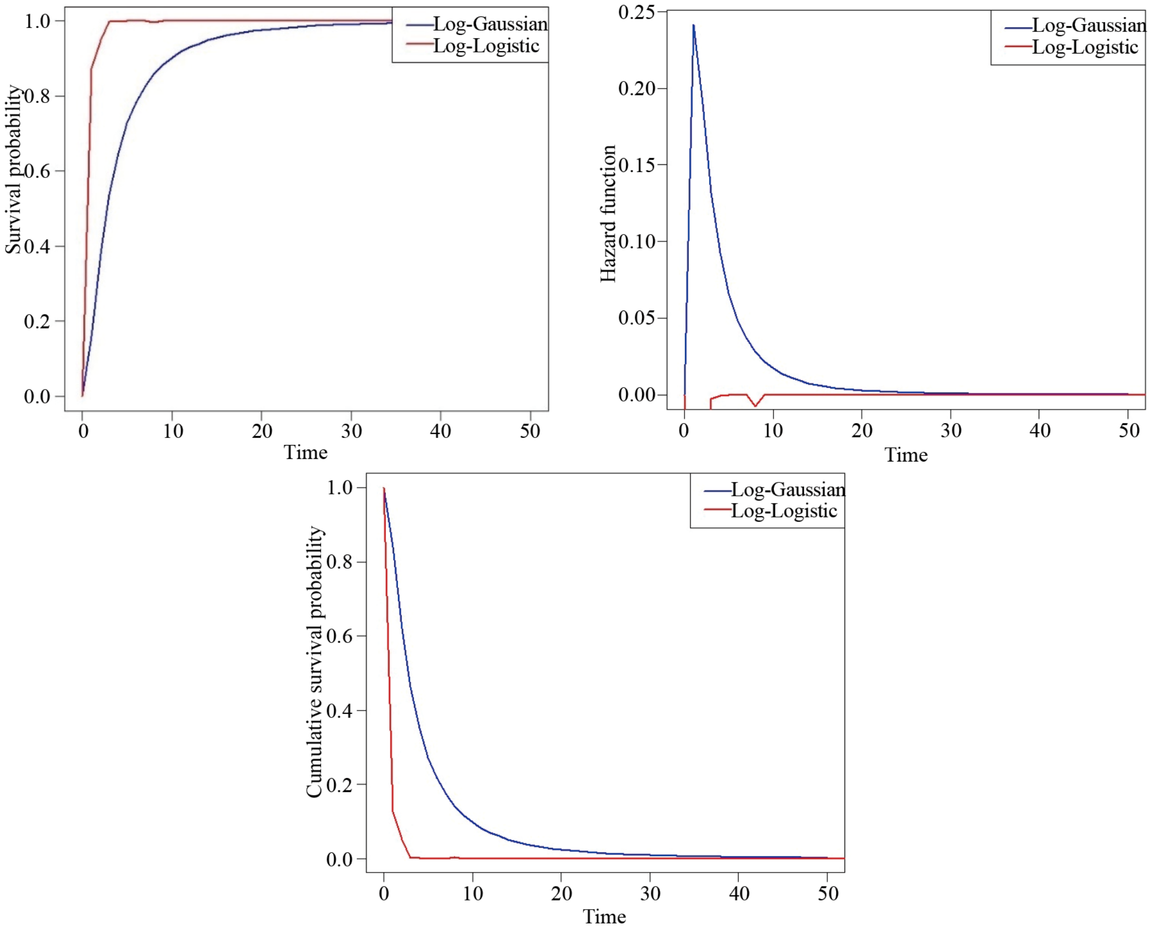

Figure 3. The sub figures represents Computed SF, HF, and CSF of LG and LL

Likewise, Table 7 displays the MSP, MCSP, and estimated AIC and BIC values as a summary of outcomes of the LG and LL regression survival models that are examined by using equations (15-18). For this assessment, a thorough comparison of the LG and LL regression models is essential. The LL regression model is the best option as it has the fewest significant components (3), an MSP of 0.9767, an MCSP of 0.0232, and the lowest values of AIC (1,661.508) and BIC (1,634.079). According to these standards, the logistic model is the most useful model for this particular dataset as it provides the optimum trade-off between model fit and parsimony.

Table 6. Evaluated metrics scores for Gaussian and Logistic

Regression models	Significant factors	MSP	MCSP	AIC	BIC
Gaussian	2	0.731971	0.2680286	1,694.025	1,666.59
Logistic	2	1	0	1,692.323	1,664.88

Table 7. Evaluated metrics scores for LG and LL

Regression models	Significant factors	MSP	MCSP	AIC	BIC
log-gaussian	3	0.9579	0.0421	1,676.139	1,648.704
log-logistic	3	0.9767	0.0232	1,661.508	1,634.079

The results of Tables 6 and 7 support the finding that the survival models based on logistic and LL regression provide the best results. The conclusion drawn from these results is that the best approach for improving comparison and guiding choices in the field of survival analysis, is to use a family of LR survival models.

Combining the Cox PH and family of Gaussian and LR survival models with advanced-survival lung cancer has gained a lot of traction in a number of fields. In this inquiry, Cox PH model is used to find a correlation between time and other pertinent factors connected to lung cancer. Time and status have been discovered to have a weakly negative association (age, ph.ecog, pat.karno). A low positive connection suggests that Ph. Karno, meal.cal, weight loss, and individuals' gender have little relationship with time. This could indicate the medical professionals that the patient's Karnofsky performance score has been improving over time; that mealtime caloric intake has somewhat increased; and that the patient has lost weight during the past six months. When survival times are positively skewed, the Gaussian model is more suitable especially when the data indicate that the survival times exhibit a symmetric, bell-shaped curve when log-transformed. When the goal is to evaluate the impact of covariates on the probabilities of survival time, a logistic model is especially helpful when the outcome is binary (e.g., survived/dead). Positively skewed survival times are modeled using both the LG and LL regression survival models, although their underlying distributions are different. The particulars of the survival data being examined will determine which of these models is best. In the end, important factors like MSP, MCSP, AIC, and BIC validated the family of LR survival models. It has been demonstrated that the suggested approach is the most effective technique for examining the lung cancer dataset. The inquiry into this topic enables us to highlight the model's effectiveness in defining survival outcomes for patients with advanced-stage lung cancer.

6. Conclusion

Regression models are crucial tools for analyzing survival data that includes variates and covariates. This study reviews four popular regression survival models, viz., the Gaussian, Logistic, LG, and LL models, which can be utilized to model survival phenomena. Upon thorough examination of the outcomes from the Gaussian and Logistic survival models, it is apparent to note the Logistic model outperforms the Gaussian models across multiple assessment criteria.

Specifically, the Logistic model exhibits significantly greater MSP values of 1, indicating a more optimistic outlook for survival outcomes. As a result, the p-values show that it contains two significant components, indicating a strong predictive capacity. Further investigation reveals that the logistic model has the lowest AIC and BIC values of 1,692.323 and 1,664.88 respectively when compared to the Gaussian model values 1,694.025 and 1,666.59.

Relatively, it is evident from a thorough examination of the data from the LG and LL survival methods that LL model outperforms the LG models on MSP values of 0.9767 and MCSP of 0.0232. The lowest AIC and BIC values are 1,634.079 and 1,661.508, respectively, and the p-values indicate that there are three significant components. Once all pertinent components, MSP, MCSP, and AIC and BIC values have been thoroughly analyzed, the family of LR survival models is the best and most efficient way to evaluate the lung cancer dataset, in view of the above results. It can be the preferred choice for researchers and professionals seeking a deep understanding of the principles of survival time in this field due to its exceptional performance on numerous dimensions.

Acknowledgement

I express my profound gratitude to the administrative authority of Vellore Institute of Technology and the School of Advanced Sciences for providing us with all the heartfelt support and exhaustive materials.

Conflict of interest

The authors attest to have no conflicting financial or other motivations to declare in relation to the present study.

References

[1] Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B, et al. Lung cancer immunotherapy: progress, pitfalls, and promises. *Molecular Cancer*. 2023; 22(1): 40.

[2] Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. *Nature Reviews Clinical Oncology*. 2023; 20(9): 624-639.

[3] Yang Q, Zhang P, Wu R, Lu K, Zhou H. Identifying the best marker combination in CEA, CA125, CY211, NSE, and SCC for lung cancer screening by combining ROC curve and logistic regression analyses: is it feasible? *Disease Markers*. 2018; 2018(1): 2082840.

[4] Li Y, Wang G, Li M, Li J, Shi L, Li J. Application of CT images in the diagnosis of lung cancer based on finite mixed model. *Saudi Journal of Biological Sciences*. 2020; 27(4): 1073-1079.

[5] Vedaraj M, Anita CS, Muralidhar A, Lavanya V, Balasaranya K, Jagadeesan P. Early prediction of lung cancer using gaussian naive bayes classification algorithm. *International Journal of Intelligent Systems and Applications in Engineering*. 2023; 11(6s): 838-848.

[6] Tirzïte M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. *Journal of Breath Research*. 2018; 13(1): 016006.

[7] Qun EL, Abd Rahman H. Prediction of lung cancer using logistics regression. *Proceedings of Science and Mathematics*. 2023; 16: 86-97.

[8] Muse AH, Mwalili S, Ngesa O, Almalki SJ, Abd-Elmougod GA. Bayesian and classical inference for the generalized Log-Logistic distribution with applications to survival data. *Computational Intelligence and Neuroscience*. 2021; 2021(1): 5820435.

[9] Lai MC, Arasan J. Single covariate log-logistic model adequacy with right and interval censored data. *Journal of Quality Measurement and Analysis*. 2019; 16(2): 131-140.

[10] Escudero-Vilaplana V, Collado-Borrell R, De Castro J, Insa A, Martínez A, Fernández E, et al. Cost-effectiveness of adjuvant atezolizumab versus best supportive care in the treatment of patients with resectable early-stage non-small cell lung cancer and overexpression of PD-L1. *Journal of Medical Economics*. 2023; 26(1): 445-453.

[11] Traore S, Sashegyi A, Winfree KB, Taipale KL, Jen MH. Bayesian survival extrapolation for cost-effectiveness analysis: a case study of RELAY for ramucirumab in combination with erlotinib in the treatment of non-small-cell lung cancer. *Journal of Medical Economics*. 2023; 26(1): 1479-1488.

[12] Wong MS, Pons A, De Sousa P, Proli C, Jordan S, Begum S, et al. Determining the optimal time to report mortality after lobectomy for lung cancer: An analysis of the time-varying risk of death. *JTCVS Open*. 2023; 16: 931-937.

[13] Ahmed LA. Parametric models in survival analysis for lung cancer patients. *Ibn AL-Haitham Journal For Pure and Applied Sciences*. 2021; 34(2): 108-118.

[14] Chan HF, Hsu WH, Chen JP, Lee JH. Factors associated with survival of patients with advanced lung cancer and long travel distances. *Journal of the Formosan Medical Association*. 2024; 123(2): 273-282.

[15] Amicizia D, Piazza MF, Marchini F, Astengo M, Grammatico F, Battaglini A, et al. Systematic review of lung cancer screening: advancements and strategies for implementation. *Healthcare*. 2023; 11(14): 2085.

[16] Sultana J, Jilani AK. Predicting breast cancer using logistic regression and multi-class classifiers. *International Journal of Engineering Technology*. 2018; 7(4.20): 22-26.

[17] Zhu H, You X, Liu S. Multiple ant colony optimization based on pearson correlation coefficient. *IEEE Access*. 2019; 7: 61628-61638.

[18] Jamil SA, Abdullah MA, Kek SL, Olaniran OR, Amran SE. Simulation of parametric model towards the fixed covariate of right censored lung cancer data. *Journal of Physics: Conference Series*. 2017; 890(1): 012172.

[19] Bhattacharjee A, Dey J, Kumari P. A combined iterative sure independence screening and Cox proportional hazard model for extracting and analyzing prognostic biomarkers of adenocarcinoma lung cancer. *Healthcare Analytics*. 2022; 2: 100108.

[20] Aako OL, Adewara JA, Are SO. Risk factor analysis of breast cancer patients in a Nigerian tertiary hospital. *FUDMA Journal of Sciences*. 2022; 6(3): 95-99.

[21] Musa FN, Bello BM, Hassan IA. Comparing the performance of a survival models in a computerized dataset. *ternational Journal of Scientific Advances*. 2023; 4(3): 354-358.

[22] Taketomi N, Yamamoto K, Chesneau C, Emura T. Parametric distributions for survival and reliability analyses, a review and historical sketch. *Mathematics*. 2022; 10(20): 3907.

[23] Amran SE, Abdullah MA, Kek SL, Jamil SA. Analysis of survival in breast cancer patients by using different parametric models. *Journal of Physics: Conference Series*. 2017; 890(1): 012169.