BIO PHARMACEUTICAL SCIENCE AND RESEARCH JOURNAL ISSN: 3049-2599

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

ON THE ESTIMATION OF SURVIVAL TIME OF CANCER PATIENTS

Malli Puttaiah.

Sri Padmathi Scooh of Pharmacy, Thirupathi.

ABSTRACT:

This study aims to evaluate several approaches for estimating survival functions in lung cancer patients using non-parametric, semi-parametric, and parametric techniques. The performance of these models is compared, and the expected survival times derived from each method are examined. Special attention is given to determining the most suitable model for the dataset analyzed.

Keywords: Survival analysis, lung cancer, Kaplan–Meier estimator, Cox proportional hazards model.

INTRODUCTION:

Survival analysis commonly employs non-parametric, semi-parametric, and parametric techniques to model time-to-event data. The Kaplan–Meier estimator, a widely used non-parametric tool, provides a simple means of estimating survival probabilities without assuming any specific underlying distribution. The Cox proportional hazards (PH) model extends this by incorporating covariates while leaving the hazard baseline unspecified, making it an essential semi-parametric method in clinical studies.

Parametric survival models offer an alternative when the underlying distribution is assumed to follow a known form, such as exponential, Weibull, gamma, or log-normal. These approaches often produce smoother survival curves and allow direct estimation of hazard and cumulative hazard functions. In this work, we focus on exponential and Lindley distributions due to their analytical simplicity and ability to represent positively skewed survival patterns.

Lung cancer remains a major cause of mortality worldwide, and understanding survival trends is crucial for planning treatment strategies and public health interventions. Previous studies have examined the influence of variables such as age, sex, tumor characteristics, performance scores, and treatment response on survival outcomes. The present study evaluates various survival models to determine those best suited for describing the survival pattern of lung cancer patients using R statistical software

NON-PARAMETRIC METHOD: KAPLAN–MEIER ESTIMATOR

The Kaplan–Meier (KM) method also called the product-limit estimator is commonly used as an initial step in survival analysis due to its minimal assumptions. The method estimates survival probabilities at observed event times and can appropriately handle right-censored data.

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

The Kaplan–Meier survival estimator is defined as

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

where:

t_i = event time,

d_i = number of events at time t_i,

n_i = number of individuals at risk just prior to t_i

In situations where the exact event time cannot be determined, KM provides a reliable approximation. It produces a step-wise survival curve, with each drop reflecting the occurrence of an event.

SEMI-PARAMETRIC METHOD: COX PROPORTIONAL HAZARDS MODEL

The Cox PH model is the most widely applied regression method in survival research. It relates the hazard rate to covariates without assuming any specific baseline hazard function.

The model is expressed as:

$$h(t|X_i) = h_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

Here,

h (t) is the hazard at time t,

h_0 (t) is the baseline hazard,

X_i represents covariates,

β values quantify the impact of each covariate.

The hazard ratio (HR), computed as (β), indicates how the hazard changes with each unit change in the covariate.

PARAMETRIC METHODS

Parametric survival models assume a specific probability distribution for survival times.

Exponential Distribution

The exponential distribution is one of the simplest lifetime distributions and assumes a constant hazard rate.

Probability density function:

$F(t) = \lambda e^{-\lambda t}, t \ge 0$

Survival function:

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

Hazard function:

$h(t) = \lambda$

Cumulative hazard

$H(t) = \lambda t$

Lindley Distribution

The Lindley distribution is widely used to model skewed lifetime data and has been applied in disciplines such as medicine and engineering.

Probability density function:

$$f(t;\theta) = \frac{\theta^2}{(\theta+1)}(1+t)e^{-\theta t},\ t > 0, \theta > 0$$

Cumulative distribution:

$$F(t) = 1 - \frac{e^{-\theta t}(1+\theta+\theta t)}{1+\theta}$$

Survival function:

$$S(t) = \frac{e^{-\theta t}(\theta t+\theta+1)}{1+\theta}$$

This distribution can capture positively skewed survival patterns, making it suitable for datasets

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

where event times cluster near the lower range and gradually taper off.

ANALYSIS OF LUNG CANCER SURVIVAL DATA

A dataset from the North Central Cancer Treatment Group, originally containing 228 observations, was reduced to 166 complete cases for analysis. Clinical and demographic variables recorded included survival time, event status, age, sex, ECOG and Karnofsky performance scores, caloric intake, and weight loss in the previous six months

Kaplan–Meier Estimates

Using R, survival probabilities and their 95% confidence intervals were calculated. The median survival time was approximately 320 days, with a one-year survival probability of about 42%. The KM survival curve showed the typical stepwise decline characteristic of lung cancer survival data.

Cox Proportional Hazards Model

Effect of Sex

Analysis indicated that females had a significantly lower risk of death than males. The hazard ratio for sex was approximately 0.62, suggesting that females had a 38% reduction in hazard compared with males.

Multivariable Cox Model

Significant predictors at the 1% level included:

- Sex
- ECOG performance score (ph_ecog)
- Physician-rated Karnofsky score (ph_karno)

Among these variables, ph_ecog showed the strongest association with hazard, indicating that performance status is a critical determinant of survival.

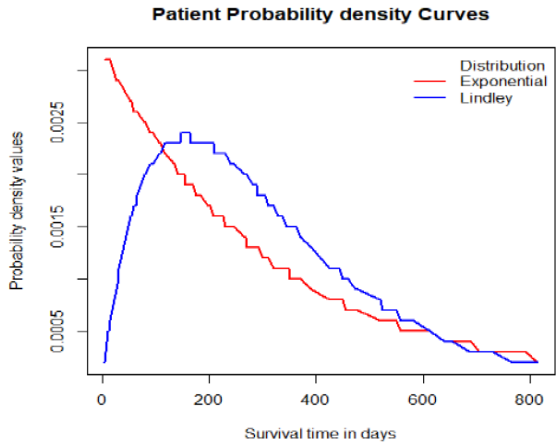
Parametric Model Comparison

Both exponential and Lindley distributions were fitted using maximum likelihood estimation in R.

Model comparison metrics:

Model	Parameter Estimate	S.E.	-2LL	AIC
Exponential	λ =0.0032	0.0003	2238	2240
Lindley	θ =0.0064	0.0004	2206	2208

Fig 1:



The Lindley distribution achieved lower -2LL and AIC values, indicating a superior fit. It also produced a larger median survival estimate (267 days) compared with the exponential model (110 days). Graphical evaluations of density and survival curves supported these findings, with the Lindley model better capturing the positively skewed distribution of survival times.

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

ON THE ESTIMATION OF SURVIVAL TIME OF CANCER PATIENTS

AIM AND OBJECTIVES

Aim

To evaluate and compare non-parametric, semi-parametric, and parametric survival models for estimating the survival time of lung cancer patients, and to identify the model that provides the best fit for the observed data.

Objectives

To estimate the survival probability and median survival time of lung cancer patients using the Kaplan–Meier method.

To assess the influence of demographic and clinical variables on patient survival using the Cox proportional hazards model.

To fit and compare parametric survival models, specifically the exponential and Lindley distributions, using maximum likelihood estimation.

To determine the most appropriate model for predicting survival patterns through algebraic and graphical evaluations (AIC, log-likelihood, survival and density curves).

To interpret the survival characteristics of lung cancer patients and identify factors associated with increased or decreased hazard rates.

MATERIAL AND METHODS:

Study Design and Data Source

This study is based on secondary data obtained from the North Central Cancer Treatment Group (NCCTG) lung cancer dataset. The dataset originally comprised 228 patients diagnosed with lung cancer. After removing incomplete and missing records, 166 patients were retained for analysis. The study follows an observational design aimed at exploring survival patterns and identifying factors influencing patient outcomes

Study Variables

The dataset includes demographic and clinical variables commonly used in survival research:

Survival time (days): Duration from study entry until death or censoring.

Censoring status:

- 1 = Alive (censored)
- 2 = Death (event)

Age: Age of the patient in years.

Gender:

- 1 = Male
- 2 = Female

ECOG performance score (ph.ecog):

- 0 = Asymptomatic
- 1 = Symptomatic but ambulatory
- 2 = In bed < 50% of the day
- 3 = In bed > 50% of the day
- 4 = Completely bedbound

Karnofsky performance score (physician-rated) (ph.karno): Ranging from 0 (poor) to 100 (excellent).

Karnofsky performance score (patient-rated) (pat.karno): Self-rated score by the patient.

Meal calories (meal.cal): Average daily caloric intake.

Weight loss (wt.loss): Weight lost in the previous six months (in pounds).

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

Statistical Methods

1. Non-Parametric Survival Analysis

The Kaplan–Meier (KM) estimator was used to compute survival probabilities at different time points and to estimate median survival time. The method accounts for right-censored observations. The KM survival curve and 95% confidence intervals were generated using the R statistical software.

2. Semi-Parametric Model: Cox Proportional Hazards (PH) Model

The Cox PH model was employed to evaluate the relationship between survival time and covariates such as age, gender, performance scores, caloric intake, and weight loss. Hazard ratios (HR) and corresponding 95% confidence intervals were estimated. Statistical significance was determined using Wald, likelihood ratio, and log-rank tests.

3. Parametric Survival Modelling

Two parametric models were fitted to the dataset:

- Exponential distribution
- Lindley distribution

Model parameters were estimated using the Maximum Likelihood Estimation (MLE) method implemented in R. For each model, the following metrics were computed:

- Parameter estimates and standard errors
- Negative log-likelihood (–2LL)
- Akaike Information Criterion (AIC)
- Probability density and survival function values

Graphical plots, including survival curves and density plots, were used to visually assess model fit.

4. Model Comparison and Validation

Model performance was evaluated using a combination of:

- AIC values (lower values indicate better fit)
- –2 log-likelihood
- Shape of fitted density and survival curves
- Median survival estimates

These criteria were applied to determine the most appropriate model for describing the survival pattern of lung cancer patients.

Software Used

All analyses were conducted using R (R-Software), utilizing packages suitable for survival analysis, including survival and custom routines for parametric model fitting.

RESULTS:

1. Kaplan–Meier Survival Estimates

The Kaplan–Meier method was applied to estimate the probability of survival for lung cancer patients. The step-wise survival curve demonstrated a gradual decline over time, reflecting the occurrence of multiple event times throughout the follow-up period.

The median survival time was estimated at 320 days, with a 95% confidence interval of 285–390 days. The probability of surviving the first year after diagnosis was approximately 42%,

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

indicating a substantial drop in survival within the initial year. The KM curve, along with its upper and lower confidence bounds, highlights the overall survival pattern and accounts for censored observations.

2. Cox Proportional Hazards Model

2.1 Effect of Gender (Univariate Model)

The univariate Cox model revealed that gender significantly influenced survival. The estimated hazard ratio for females compared with males was 0.62, suggesting that female patients had a 38% lower risk of death during the observation period. The effect was statistically significant ($p = 0.015$), indicating a notable difference in survival outcomes between the two groups.

2.2 Multivariable Cox Model

When multiple covariates were included in the Cox model, the following variables showed statistical significance:

- Gender (HR < 1): Females had better survival outcomes.
- ECOG performance status (ph.ecog): Higher ECOG scores were strongly associated with increased hazard, making this the most influential predictor of mortality.
- Physician-rated Karnofsky score (ph.karno): Higher scores indicated better functional status and were associated with reduced hazard.

Other variables such as age, patient-rated Karnofsky score, caloric intake, and weight loss did not reach statistical significance in the multivariable context. Overall model fit indices—including concordance (0.651) and likelihood ratio tests—confirmed the suitability of the Cox PH framework for explaining variability in survival times.

3. Parametric Model Fitting

3.1 Exponential vs. Lindley Models

Two parametric survival models were fitted using maximum likelihood estimation. Table 4 results showed:

- The Lindley distribution produced smaller –2 log-likelihood (–2LL = 2206) and AIC (2208) values than the exponential model (–2LL = 2238; AIC = 2240).
- The median survival estimates were:
    - Exponential model: 110 days
    - Lindley model: 267 days

The graphical comparison demonstrated that the survival and density curves from the Lindley model more accurately captured the skewed structure of the data, where many events occurred early, and fewer occurred at later times. The exponential model, assuming a constant hazard, provided a comparatively poor representation of the observed pattern.

3.2 Density and Survival Plots

The density plots illustrated that survival times were positively skewed, with most observations concentrated between 0 and 200 days. The tail gradually tapered to the right, supporting the

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

suitability of the Lindley distribution for this dataset. The Lindley survival function also aligned more closely with the empirical KM curve, offering a better predictive fit.

DISCUSSION

This study evaluated different survival modelling approaches—non-parametric, semi-parametric, and parametric—to characterize the survival experience of lung cancer patients. The findings highlight the strengths and limitations of each approach and offer insights into factors influencing patient outcomes.

The Kaplan–Meier estimator provided a clear and assumption-free summary of the survival pattern. The median survival of 320 days is consistent with established evidence that lung cancer has a relatively poor prognosis, with a steep decline in survival probabilities during the first year. The KM curve effectively described the overall trend while accommodating censored cases.

The Cox PH analysis revealed that performance status plays a central role in determining survival. Specifically, ECOG performance score (ph.ecog) emerged as the strongest predictor, indicating that patients with diminished functional ability face a significantly increased risk of mortality. This aligns with clinical expectations, as poor performance status often reflects advanced disease stage and reduced tolerance to therapy. Gender also demonstrated a significant association with survival, with females showing better outcomes than males a finding frequently reported in lung cancer literature and possibly linked to biological and behavioral factors.

The comparison of parametric models provided additional insights into the shape and distribution of survival times. The lung cancer dataset exhibited positive skewness, with a concentration of events occurring in the earlier months following diagnosis. In this context, the Lindley distribution offered a superior fit compared with the exponential model.

Its flexibility in representing skewed data allowed it to capture the empirical survival pattern more accurately. The exponential model, relying on a constant hazard assumption, was less effective, reflected by higher AIC and poorer alignment with the observed survival curve.

Taken together, the results demonstrate that the Lindley distribution, supported by both graphical and statistical criteria, is the most appropriate model for this dataset. Additionally, the study reinforces the usefulness of combining non-parametric, semi-parametric, and parametric methods to obtain a comprehensive understanding of survival characteristics in clinical populations

CONCLUSION

This study examined survival patterns among 166 lung cancer patients using a combination of statistical models. Approximately 28% of the observations were censored, with

DOI: https://doi.org/10.63712/bpsrj-v1i4p001

males representing 62% of the sample. The mean age was 62.6 ± 9.2 years.

Key findings include:

- The Kaplan–Meier method estimated a median survival of 320 days and a first-year survival probability of around 41–42%.
- Cox regression results showed that females had significantly better survival than males, and ECOG performance status was the strongest predictor of mortality.
- Among the parametric models assessed, the Lindley distribution provided a notably better fit than the exponential distribution and produced higher survival time estimates.

Overall, both graphical and algebraic evaluations indicate that the Lindley distribution is more appropriate for modeling lung cancer survival data

REFERENCE:

1. Forde PM, Spicer J, Lu S, et al. Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer. N Engl J Med. 2022; 386(21):1973-1985.
2. Jing-Yang Huang, Chuck Lin, Stella Chin-Shaw Tsai and Frank Cheau-Feng Lin (2022). Human Papillomavirus Is Associated with Adenocarcinoma of Lung: A Population-Based Cohort Study. Frontiers in Medicine 2022; 9:1-11.
3. Julie A. Barta, Charles A. Powell and Juan P. Wisnivesky (2019). Global Epidemiology of Lung Cancer. Annals of Global Health 2019; 85(1): 8, 1–16.
4. Lawless JF (2003) Statistical Models and Methods for Lifetime Data (2nd edn). John Wiley & Sons, Inc: New Jersey.
5. Miller RG Jr (1981) Survival Analysis. John Wiley & Sons: New York.
6. Satar R, Ali A, Abraha W, et al (2016). Estimating the economic burden of lung cancer in Iran. Asian Pac J Cancer Prev, 17, 4729.
7. Siegel RL, Miller KD, Wagle NS, Jemal A (2023). Cancer statistics, 2023. CA Cancer J Clin. 2023;73(1):17-48.
8. C.Zirafa, Gaetano Romano, Elisa Sicolo, Andrea Castaldi, Federico Davini, Franca Melf (2023).

Source of Support: Nil. **Conflicts of Interest:** None