Malaysian Journal of Mathematical Sciences 20(2): 723–737 (2026)
https://doi.org/10.47836/mjms.20.2.14

Malaysian Journal of Mathematical Sciences

Journal homepage: https://mjms.upm.edu.my

# Gender Differences in Lung Cancer Survival Based on Non-parametric and Parametric Methods

Shankaraiah, P.[1], Mokesh Rayalu, G.*[2], and Prasad, A. V.[3]

[1]*Department of Mathematics, SJC Institute of Technology, Chickballapur-562101, India*
[2]*School of Advanced Sciences, Vellore Institute of Technology, Vellore, 632014, India*
[3]*National Assessment and Accreditation Council, Bengaluru, 560072, India*

*E-mail: mokesh.g@vit.ac.in*
**Corresponding author*

*Received: 1 February 2025*
*Accepted: 15 October 2025*

## Abstract

Non-parametric and semi-parametric models allow us to predict the gender of individuals based on their survival time in lung-survival data. Primary, we used the North Central Cancer Treatment Group for patients with advanced lung survival dataset to apply the combination of survival models, along with the relevant confidence intervals. Based on the findings of this research, gender is statistically significant based on $p-$value, determinants in predicting coefficient of $-0.5509$, and hazard ratio of $0.5765$ using cox proportional hazard with Breslow model. The results of these methods' analysis, $\chi^2$ test, were utilized to ascertain whether a gender difference exists. The Kaplan-Meier curve for the entire dataset estimates a mean survival time of $327.5$ days and a median survival time of $310$ days; for males, the estimated median overall survival is $270$ days, and for females, the median survival time is estimated to be $426$ days. A study of the dataset indicates that sex significantly influences survival results. Gender is a determinant that impacts the duration of survival, with females often exhibiting a higher median survival time of $426$ days. These findings emphasize the importance of considering age and gender when predicting survival outcomes and creating targeted interventions for certain subgroups.

**Keywords:** central tendency values; lung cancer; probability survival time; survival metrics; survival models.

## Abbreviation's

| | | |
|---|---|---|
| CIs | : | Confidence intervals |
| Cox PH | : | Cox proportional hazard |
| HR | : | hazard ratio |
| KM | : | Kaplan-Meier |
| LC | : | Lung cancer |
| MDST | : | Median survival time |
| MST | : | Mean survival time |
| OS | : | Overall survival |

## 1 Introduction

Lung cancer (LC) constitutes a significant universal health care issue, ranking as the foremost cause of cancer-related mortality [29]. Its high prevalence and advanced diagnosis stage make it dangerous, reducing treatment options and impacting survival rates [30]. Factors like tobacco smoke, environmental contaminants, workplace risks, and inherited tendencies contribute to its complexity [11]. Unlike many other types of cancer, LC is categorized into non-small cell and small cell types [20]. The prognosis for later stages of LC remains generally gloomy, emphasising the necessity of finding creative approaches to research and management [16]. To enhance patient results, new approaches to early identification, targeted treatments, and personalised medicine are required [5]. LC is the most dangerous cancer globally, more than $1.79$ million new cancer cases are expected to be diagnosed in $2020$. Colon cancer is the second most lethal type of cancer, with stomach and liver cancer coming in second [26]. But heart disease and stroke account for the largest percentage of deaths globally, making LC only the sixth most common cause of death [9]. Globally, LC is the primary cause of cancer-related deaths. In $2023$, it is projected that 238,340 people will receive a LC diagnosis. In one's lifetime, $1$ in $16$ men and $1$ in $17$ women will receive a LC diagnosis [4]. Since $2006$, the incidence of LC has declined by $2.6\%$ for males and $1.1\%$ for women annually, most likely as a result of improved therapies and early detection [17].

The main objectives of this work are

- The objective is to analyze multiple variables in the dataset and determine their impact on survival time, specifically focusing on a coefficient between time, age, and sex.
- In order to determine the significant factors based on $p-$values, standard error, hazard ratio (HR) values applying the semi parametric model.
- Next, to evaluate the influence of gender on LC survival outcomes. This can be achieved by using the log-rank test and thereafter determining the survival probability.
- Lastly, to find the survival time and percentage of deaths related to sex using the Kaplan-Meier (KM) estimator.

The research provides essential information for assessing survival analysis screening using publicly available LC data. It highlights the importance of studying LC globally and analyzed semi-parametric and non-parametric estimation techniques. The rest of this article is organized as follows: The works by different authors relevant to LC analysis are described in Section 2. In Section 3.1, the data set information is used. The implemented methodologies and the metrics used for experimentation are discussed in Section 3.2. The obtained findings and implications are shown in Section 4. The conclusion has been placed subsequently in Section 5.

## 2 Related Works

The study investigated by Chaudhry et al. [8], the distinction between actuarial and conditional survival after lung resection. Using the KM, Cox Proportional Hazard (PH), and log-rank approaches, it comprised 863 patients who had anatomic lung resection between 2010 and 2021, over a median surveillance period of 44.1 months. The findings revealed that at every stage following surgery, conditional overall survival (OS) and disease control survival exceeded actuarial rates. Disease-free survival was correlated with variables such male sex, tumor size, node positivity, and the American Joint Committee on Cancer stage. According to the study, Alifano et al.[1] analysed the LC patients who are considered high-risk candidates for perioperative targeted and immunological therapies may benefit from prognostic assessment, which is essential for cancer treatment planning. With an emphasis on OS up to five years, the study examined clinical, surgical, and pathological data from ongoing patients with non-small-cell LC between 2003 and 2020. It examined nine prognostic markers and their effect on nomograms in $62,633$ individuals with a median survival period of 9.2 years. Using sex-specific nomograms and Cox modeling, prognostic relevance was evaluated.

Jiang et al. [14] comparative analysis of survival outcomes in advanced NSCLC patients engaged in clinical trials versus those who were not shed light on the transformative potential of clinical engagement. Their robust analysis underscored the marked impact of clinical trial participation of 155 patients on both OS and progression-free survival by using KM curves and the Cox PH model, elucidating the pivotal role of trial involvement in advancing treatment paradigms and improving patient outcomes. An independent UK investigation, Teng et al. [28] discovered that ceasing smoking upon diagnosis is strongly linked to enhanced survival in non-small-cell LC, irrespective of the stage. The study monitored a cohort of 2751 individuals diagnosed with non-small-cell LC for a maximum duration of two years or until they passed away. Patients were provided with the option to quit smoking in accordance with national norms and local services. The preliminary two-year KM and Cox analyses were conducted. The proportional hazard functions for individuals who quitter and those who remain were shown to be significantly dissimilar. The median survival time (MDST) for individuals who quitters were 659 days, while it was 348 days for those who continuers. This study suggests that smokers diagnosed with LC should be provided with smoking cessation guidance and therapies.

The study examined by Alvarez et al. [3] of survival disparities based on gender and screening status using the Boston LC Study, which examined 1216 patient databases, highlighted critical differences in survival rates by using KM curves and Cox PH models. Employing propensity score matching, their findings unveiled substantial disparities between screened and unscreened males, accentuating the urgency for widescale LC screening initiatives to mitigate the gender-based survival gap. The study investigated by Sun et al. [27] provided 340 patients with significant insights into the implications of genetic changes on OS, recurrence-free survival, and the effect of clinicopathological variables using the KM and Cox PH model. This was achieved by delving deeply into the genetic foundations of stage I solid-appearing NSCLC. Their thorough retrospective study identified common mutations, such as those in EGFR, KRAS, ALK, ROS1, and fusion genes, highlighting their influence on recurrence-free survival. However, given the implications for OS, more research is necessary to fully comprehend their significance in patient outcomes. concentrating on predictive modelling.

The following are the restrictions that we have found. Ignoring the correlation between intraoperative variables and survival results in order to offer a more thorough evaluation of treatment efficacy [1]. The clinical relationship between identified genes and patient outcomes was not further discussed by the authors [15]. In order to get stronger evidence of causal links, the author

recommended conducting more prospective studies and allowing for better control of confounding variables [14]. According to Ljunggren et al. [19] used KM, logistic, and Cox PH to estimate survival time, 95% Confidence Intervals (CIs), and HRs without identifying a significant factor, suggesting a preliminary study is needed. It should be noted that the investigation that was described above examined survival rates in relation to OS rather than clusters. Still, the majority of the works use KM, Cox proportions, and log rank concepts separately, with fewer combining them. The study analyzed datasets for various variables to compare gender with MDST and mean survival time (MST), survival time probability, and death percentages. The KM estimator produces survival curves, allowing comparisons by sex with respect to survival time.

## 3 Materials and Methods

### 3.1 Data set and processing

This investigation used survival statistics from the North Central Cancer Treatment Group for patients with advanced LC. The patient's activities of daily living capacity are indicated by their performance scores. The data was provided by Terry Therneau. All of these variables are features in the data sample, which consisted of 168 rows and 9 columns. The study spanned a follow-up period of 1022 days. I'm utilising all the units in this data to identify status with respect to time is the dependent and remaining all are the independent variables.

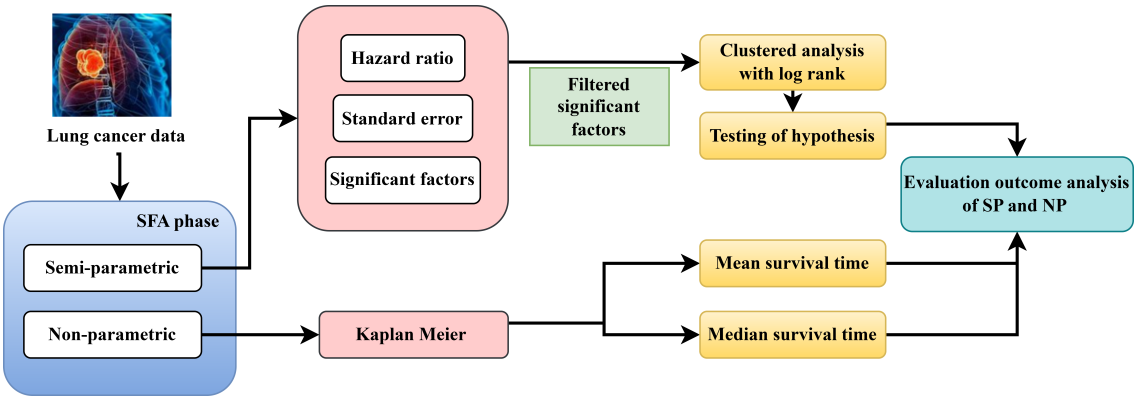


Figure 1: The proposed survival methodological architecture.

Considering the dataset generated within the context of patients suffering from advanced LC, the whole pipeline functionality for the processing of data is depicted in Figure 1. Taking into account that the dataset consists of both numerical and discrete variables [22]. The variables "Time" and "status," which indicated whether a patient was censored or died at the conclusion of the follow-up period, respectively, describe the patients' follow-up durations. To create survival times, both of these individual factors were utilized. To make it easier to look at subgroups, the people were split into three categories based on their gender (male and female) and age: less than 60 years (between 30 and 59 years), 60 years or more, and all ages. The goal of this stratification was to figure out how survival rates change between genders and between different age groups. Age, gender, performance status, Karnofsky performance score, meal calories, and weight reduction are examples of demographic and clinical characteristics [7]. While gender may have an impact on survival outcomes, age is a critical predictor in survival analysis. General well-being

and activity level are indicated by performance status and Karnofsky performance scores. Weight reduction and meal calories are examples of nutritional factors. Entire data processing [2].

The study examined the predictive variables linked to LC patients' survival duration. Based on the state of LC, researchers could utilize survival analysis, which has multiple models, particularly in identifying the significant prognostic factors using Cox PH [24]. The survival probability of patients with LC was calculated using the KM model. The significance difference between the patient's male and female groups was then ascertained using a log-rank test.

In particular, we can understand the relationships between variables according to this fundamental concept in statistics and data analysis. The covariates shown in Figure 2, exhibit a pairwise association in the majority of cases when utilizing Pearson correlation formula [31]. which, given the current circumstances, calls for a suitable choosing parameters process.

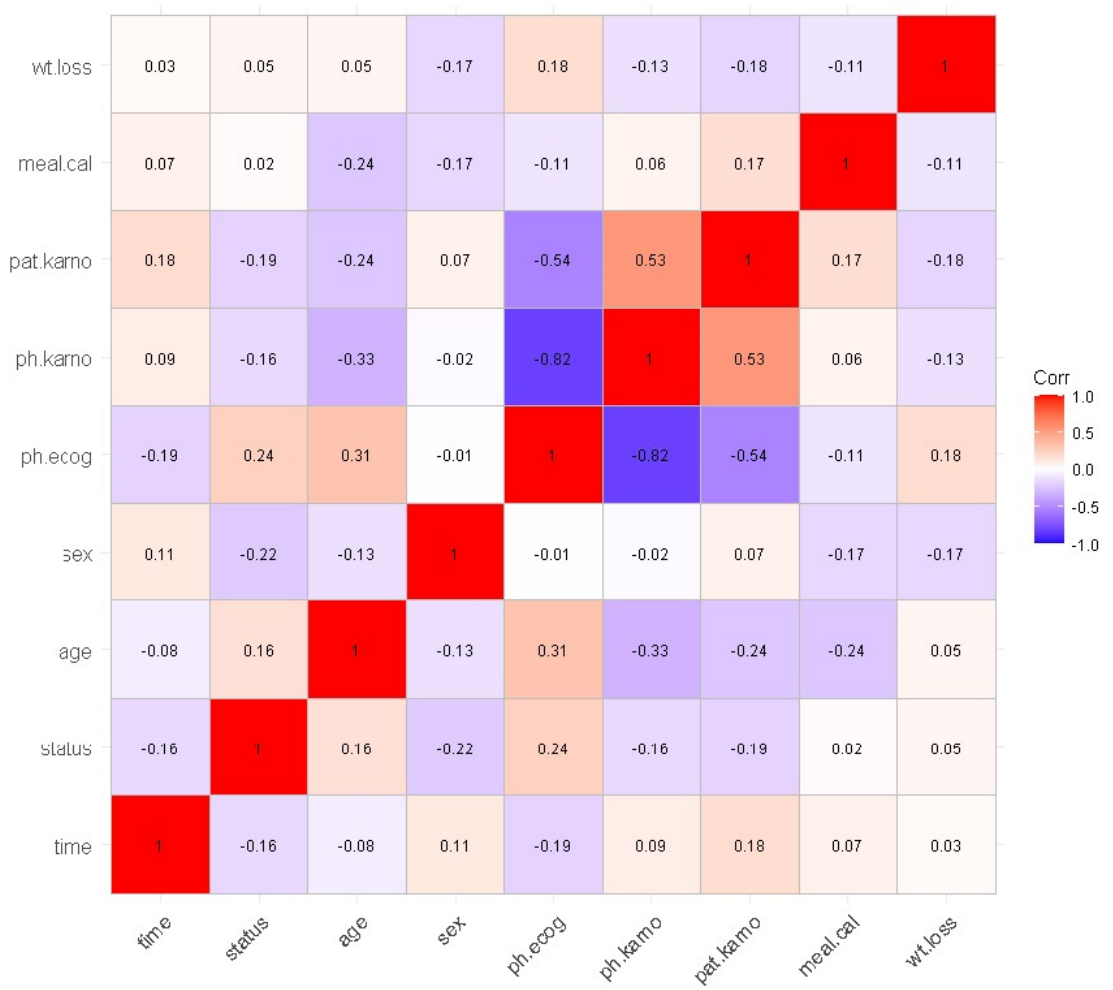


Figure 2: Pairwise analysis of correlated features in LC.

This suggests that the colour gradient shows the intensity and direction of the correlations, with deep red meaning a strong positive connection and deep blue meaning a strong negative

correlation. The correlation analysis shows that the status variable, which shows whether an event happened, is only weakly linked to individual demographic, clinical, and nutritional factors. There isn't a strong linear relationship between any of the factors and status, but there are some small patterns. For example, performance scores (Ph.karno, Pat.karno) and sex have negative correlations with status, whereas age, ECOG score, meal.cal, and wt.loss have positive connections. These results show that no one event can accurately predict the probability of death. So, a multivariate estimation of survival perform is necessary to provide a better picture of how many factors affect survival outcomes in LC patients [25].

### 3.2 Statistical analysis

Let $T$ represent the survival time $t$ as a continuous variable, $x_i = (x_{i1}, x_{i2}, \ldots, x_{i7})$ represent the observed values of the covariates with regard to subject $i$ [6]. The stratified Cox regression model's HF has the following general form,

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

The Cox model uses a partial likelihood that focuses on the regression coefficients $\beta_i$. The partial likelihood is based on the concept that the risk of every person who is at risk at time t is proportionate to the likelihood of each person failing at that moment [13]. We examined the proportional hazards assumption to make sure the Cox PH model was correct. We utilized Schoenfeld residuals to see if the effect of covariates varied over time. The assumption was true for the variables that were included because the $p-$value was not significant ($p > 0.05$). Additionally, log(log survival plots) were utilized as a graphical approach to assess parallelism between groups, which further confirmed the proportionality of hazards throughout time.

The log partial likelihood function,

$$\ell(\beta) = \sum_{j=1}^{m} \left[ \sum_{i \in D(t_j)} X_i\beta - \log \left( \sum_{k \in R(t_j)} \exp(X_i\beta) \right) \right]. \tag{2}$$

To estimate the regression coefficients $\beta_i$, the log partial likelihood $\ell(\beta)$ is maximized. The Newton–Raphson method is applied to solve the maximization problems. Once the regression coefficients $\beta_j$ are estimated, the $HR$ for each covariate is calculated as,

$$HR_j = \exp(\beta_j). \tag{3}$$

To compare the survival times of the male and female groups, a supplementary log rank test is employed as a hypothesis test [27]. The log rank test is similar to a chi-square ($\chi^2$) test.

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}, \tag{4}$$

where, $k-1$ is the degrees of freedom, $O_i$ is the total number of events observed at each time point, $E_i$ is the total number of events predicted at each time point. Calculate the $p-$value using the chi-square distribution and interpret the results based on the $p-$value and significance level.

For the Lung-Survival Dataset, the KM test is a reliable statistical approach for effective age and sex [12]. The MST and MDST for patients in two groups are calculated using this test. The KM estimator [15] can be defined as,

$$S(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right), \qquad j = 1, 2, 3, \ldots, 168, \tag{5}$$

where $r_j$ indicates number of persons at risk and $d_j$ represents persons who die at a certain time.

Standard $95\%$ CIs [23] for $S(t)$ assume the following form,

$$\hat{S}(t) \pm Z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}(t). \tag{6}$$

MST and MDST metrics are utilized in medical research to analyze survival statistics in LC data, characterizing the average length of survival time following disease diagnosis [10], like cancer.

$$\text{MST} = \frac{\sum(\text{Events time of each individual})}{\text{Number of individuals}}. \tag{7}$$

The MST may not fully represent community survival experiences, so additional statistical metrics like MDST or survival rates at specific time periods can provide more comprehensive knowledge [18]. The MDST is a crucial survival metric used in disease, epidemiology, and clinical trial settings to estimate the time it takes for $50\%$ of monitored patients to experience an event [21], aiding in prognosis, treatment planning, and therapy efficacy evaluation. It is derived from the KM estimator by utilising at $95\%$ CIs to get the MDST. The significance threshold Sex $0.00609$, Ph.ecog $0.00101$, and Ph.karno $0.04574$, and the software R version $4.2.3$ used for the analysis.

## 4 Results and Discussion

The present research study makes use of every unit in this data set which is enunciated in Table 1.

Table 1: Present the descriptive statistics of all variables, types and distributions.

| Variable | Type | Normality | Statistical measure |
|---|---|---|---|
| Age | Quantitative | Normally distributed | $[62.61 \pm 9.17]$ |
| Sex | Categorical | Non-normally distributed | $1.00$ |
| Ph.ecog | Quantitative | Non-normally distributed | $1.00$ |
| Ph.karno | Quantitative | Non-normally distributed | $80.00$ |
| Pat.karno | Quantitative | Normally distributed | $[79.58 \pm 15.01]$ |
| Meal.cal | Quantitative | Normally distributed | $[927.84 \pm 411.36]$ |
| Wt.loss | Quantitative | Normally distributed | $[9.78 \pm 13.32]$ |

The results of a Cox PH with Breslow regression model are shown in Table 2; each row corresponds to a distinct covariate (or feature) in the model. In (2) generated the covariate's estimated regression coefficients: a positive number means that the event's risk rises as the covariate does, whereas a negative value means that the event's risk falls as the covariate rises. In (3) and the associated graph in Figure 3 are used to get the HR. With all other variables held constant, it shows the relative chance of the event happening given a one-unit increase in the covariate. A higher risk is indicated if HR $> 1$, while a lower risk is indicated if HR$< 1$.

Table 2: Statistical metric values of Cox PH with the Breslow method.

| **Variable** | **Coefficient** | **S.E** | **HR** | **95% CI's** | $p-$**values** |
|---|---|---|---|---|---|
| Age | 0.01065 | 0.01161 | 1.01070 | [0.9880, 1.0340] | 0.35906 |
| Sex | −0.55090 | 0.20080 | 0.57650 | [0.3889, 0.8545] | 0.00609 |
| Ph.ecog | 0.73420 | 0.22330 | 2.08380 | [1.3452, 3.2277] | 0.00101 |
| Ph.karno | 0.02246 | 0.01124 | 1.02270 | [1.0004, 1.0455] | 0.04574 |
| Pat.karno | −0.01242 | 0.00805 | 0.98770 | [0.9722, 1.0034] | 0.12316 |
| Meal.cal | 0.00003 | 0.00002 | 1.00003 | [0.9995, 1.0005] | 0.89791 |
| Wt.loss | −0.01433 | 0.00777 | 0.98587 | [0.9709, 1.0009] | 0.06518 |

The interval gives a range of values that, with $95\%$ CIs, represent the genuine HR. The covariate could not be statistically significant if the period contains $1$. The coefficient estimate's correctness is gauged by the standard error. While a higher SE denotes greater uncertainty, a lower SE suggests a more accurate assessment.

The findings of a Cox PH with Breslow regression model, which evaluates the influence of different clinical trials and demographic factors on the quality of life for patient's outcomes, are displayed in Table 2. The coefficient, HR, standard error (SE), $95\%$ CIs, and statistical significance as shown by the $p-$value are used to assess each covariate. With HR $= 0.5765$ and a negative coefficient $(-0.5509)$, sex is an accurate metric of survival, as seen by the declining risk of the events. Sex is a statistically significant factor, as evidenced by the $p-$value $(0.00609)$ being far below the traditional cutoff of $0.05$ and the HR's CIs not including $1$. According to a $p-$value with a positive coefficient $(0.7342)$ and HR $= 2.0838$, Ph.ecog, or the ECOG performance score, is another significant predictor. This means that a higher ECOG score (worse performance status) more than doubles the risk of the event, and the $p-$value $(0.00101)$ indicates that this covariate is statistically significant. The CIs does not include $1$. This finding emphasizes the significance of performance status as a predictive indicator in survival analysis by indicating that patients with worse Ecog performance status have a significantly increased chance of unfavourable outcomes. With HR of $1.0227$ and a positive coefficient $(0.02246)$, Ph. Karno (physician-assessed Karnofsky score) likewise suggests a small increase in danger with higher Karnofsky scores. The $p-$value $(0.04574)$ is just below $0.05$, and the CIs barely excludes $1$, indicating that this covariate is borderline significant. Although the impact size is tiny and its practical significance may be restricted, this data suggests that higher Karnofsky scores, as determined by physicians, are related with a minor increase in the probability of the occurrence.

According to the analysis, the HR is $1.0107$ and the coefficient for age is positive $(0.01065)$, meaning that the risk increases by $1.07\%$ for every year that goes by. Nevertheless, the high $p-$value $(0.35906)$ and $95\%$ CIs imply that this impact is not statistically significant. Similar to this, the patient-assessed Karnofsky score (Pat.karno) indicates a modest decrease in the overall risk of an incident with higher scores, as seen by its negative coefficient $(-0.01242)$ and HR of $0.9877$. However, its $p-$value of $0.12316$ and CIs show no discernible effect. With a strong $p-$value $(0.89791)$, a CIs, and an insignificant positive coefficient $(0.00003)$ and HR $(1.0003)$, meal.cal, which represents caloric intake per meal, did not significantly impact survival outcomes. Finally, the HR of $0.98587$ and the negative coefficient $(-0.01433)$ for weight loss (Wt. loss) indicate that weight loss is linked to a marginal decrease in hazard. Its $p-$value $(0.06518)$ and CIs indicate a marginally significant effect on survival.

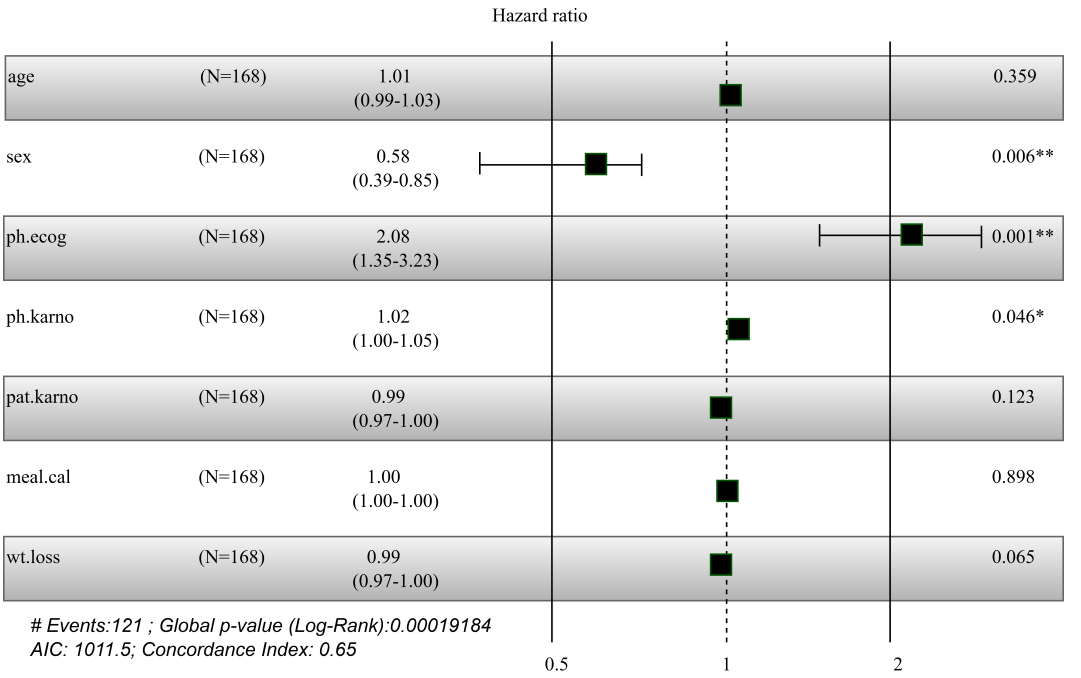


Figure 3: HR values and significant features using Cox PH model.

Figure 3 shows that the two best predictors of survival in this model are Sex and Ecog performance score, both of which have statistically significant impacts on risks. Patient-assessed Karnofsky score and calorie intake are not significant predictors, although age, weight reduction, and physician-assessed Karnofsky score have lesser, borderline effects. While other factors like age and weight loss may have more minor or non-significant effects in this particular dataset, our findings highlight the significance of sex and performance status in predicting survival outcomes.

Table 3: Gender cluster wise outcomes of a chi-square $(\chi^2)$ test.

| Gender | Observed Value | Expected Value | Formula |
|---|---|---|---|
| Males | 104 | 91.6 | 4.55 |
| Female | 64 | 73.4 | 5.68 |

Table 3 displays the outcomes of a chi-square $(\chi^2)$ test, utilized to ascertain the difference exists between the male and female groups. This test assesses whether the gender distribution in the sample significantly deviates from the expected distribution under $H_0$. The observed numbers denote the precise counts of males and females within the sample. There are $104$ males and $64$ females in this instance. The anticipated values are the frequencies that would be predicted under the null hypothesis. The anticipated count for males is $91.6$, but for females, it is $73.4$. The values are derived from the overall sample size and the anticipated proportions for each gender group. The Chi-square contribution for males is $4.55$, while for females it is $5.68$, yielding a total Chi-square statistic of $10.3$, as calculated utilizing $(4)$. This number contrasts to the crucial value of $6.632$ from the Chi-square distribution table at $1$ d.o.f and level of significance $0.01$. Consequently, we reject the null hypothesis, and the chi-square test produces strongly indicate a large disparity in gender distribution within the sample. A $p-$value of $0.001$ signifies that the observed frequencies

of males and females deviate from the expected values to an extent that is improbable to result from a random effect.

The research study conducted using (6) and (7) of the KM estimator approaches produced results that demonstrated a remarkable degree of consistency between the sexes. With a median survival duration of 270 days and a strikingly high death rate of 81.2%, the estimated 95% CI for survival for males is between 212 and 310 days. The 95% CIs for females is broader, ranging from 348 to 550 days. This shows that the median survival period for females is substantially higher at 426 days, while their death rate is lower at 58.9%. Table 4 displays these findings, while Figure 4 visually depict the various gender-specific survival attributes. Males had a higher death rate but a shorter MDST. The mortality rate for females is significantly lower than that of males, even though the former has a higher median survival duration. "More survive" is determined based on the particular criteria and analysis goals. Although suffering a lower proportion of deaths, females exhibit a higher survival outcome despite having a longer MDST.

Table 4: Gender cluster wise obtained results.

| **Sex** | 95% **CI** | **MDST** | **Percentage of deaths** |
|---|---|---|---|
| Male | $[212, 310]$ | 270 | 81.2% |
| Female | $[348, 550]$ | 426 | 58.9% |

To improve survival results for both males and females, these findings may inform customised healthcare methods, risk assessments, and the creation of gender-specific therapies. Studies surveyed a dataset and found that gender had an effect on survival rates. The fact that women fared better in terms of MDST raises questions about potential gender disparities in vulnerability, treatment efficacy, or other variables impacting longevity. Using KM survival analysis also laid the groundwork for future investigations into the reasons impacting these differences.

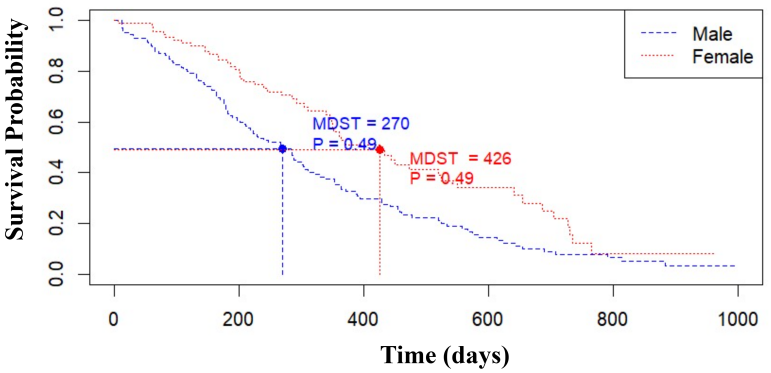


Figure 4: KM gender wise survival curve.

According to Figure 5, the overall data set's mean survival period of 327.5 days indicates that, on an average, people lived for roughly 327.5 days following a lung-related event or diagnosis. 50% of individuals survived for less than or more than 274 days. according to the MDST of 274 days. The MST is greater than the median when these two measurements are compared. This disparity can indicate that the survival times have a positively skewed distribution or that there are a few outliers. The mean can be raised by outliers, or people who lived noticeably longer than the majority, since their effects tend to be greater than those of the median.

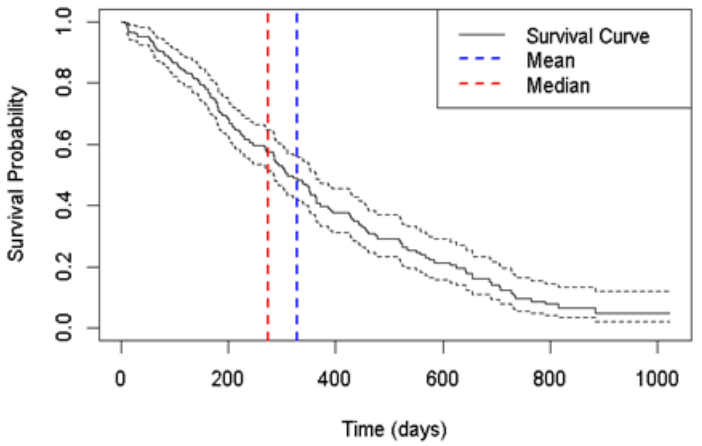


Figure 5: KM Estimator survival curve.

We have tackled the issue of possible bias related to gender and clinical factors in this work. This is a technical issue that goes apart from the typical biases related to research strategy and implementation because it is a fundamental aspect of survival model technique. Three distinct models of survival analysis were used: the KM estimator, the log-rank test, and Cox PH with Breslow. Cox PH states that sex, Ph.ecog, and Ph.karno are significant factors; they are linked to a gender of results when contrasted to complete examinations, with a $p-$value $< 0.05$, the negative coefficient value, and the decrease in the HR of the sex group to the complete evaluation group below $1$. The percentage of dangers is inversely connected with this effect on HR. Secondly, studies assessing LC survival have shown significant gender disparities in patient survival, with females consistently having a greater survival rate than males.

In Table 4, our research showed that women had a $58.9\%$ lower probability of dying from advanced LC than men. In contrast, a study that found that females had a $22.3\%$ lower relative risk of survival than males accomplished. After the dataset was divided into gender groups, it was discovered that $112$ out of $138$ males had the incident, resulting in a MDST of $270$ days, and the male groups median survival period had a $95\%$ CIs of $212$ to $310$ days. Conversely, $53$ out of $90$ females had the occurrence, which means their MDST was substantially greater $(426$ days$)$. The median survival for the female group had a wider $95\%$ CIs, ranging from $348$ to $550$ days. The observed discrepancy in survival periods between the mean and median values may indicate the existence of outliers or a skewed distribution in the dataset, which might be attributed to a subset of individuals who experienced much longer survival times than the majority. The lung survival information was thoroughly analyzed, revealing differences in survival durations between male and female categories. It raises the possibility of demographic variations influencing how patients with lung-related disorders respond to treatment, how the disease progresses, or other aspects that affect survival rates.

## 5 Conclusions

The Cox PH with Breslow model result indicates that, according to the $p-$value, Sex and Ecog performance score are the best predictive factors of survival in this model, with statistically sig-

nificant impacts on risks. Patient-assessed Karnofsky score and calorie intake are not significant predictors, although age, weight reduction, and physician-assessed Karnofsky score have lesser, borderline effects. While other factors like age and weight loss may have more minor or non-significant effects in this particular dataset, our findings highlight the significance of sex and performance status in predicting survival outcomes. Additionally, the results of the chi-square test clearly indicate that the gender distribution of the sample differs significantly.

Given the $p-$value of $0.001$, it seems improbable that the observed frequencies of males and females are the result of random variation because they deviate so much from the expected values. This leads to the rejection of the null hypothesis, suggests that there is a substantial difference between the gender distribution in this dataset and what would be predicted under equal or proportional representation. This implies a statistically significant divergence between the expected and observed gender frequencies. The KM plot, which displays a distinct difference in survival probability between males and females, there were notable differences in the MDST of the dataset between gender groups. The MDST for the entire dataset was found to be $310$ days. However, when gender was taken into account, women demonstrated the longest median survival duration of $426$ days, which was much longer than the MDST of $270$ days for men.

For the particular demographic groups under investigation, a greater MDST in this dataset seems to imply a longer survival length. The results point to possible variations in survival rates between male and female groups. In general, When the survival time increases, the survival probability rate likewise increases, and there is a subsequent decrease in the death percentage number. Concludes by showing that women have significantly higher survival rates and a lower risk of dying from advanced LC than men. With regard to these estimations, the statistical uncertainty is shown as the CIs. Furthermore, our research did not uncover statistically significant impacts of age, weight change, or daily calorie intake on the survival of patients with advanced LC, despite our exploration of these components' effects.

**Acknowledgement** We would like to convey our sincere gratitude to Vellore Institute of Technology and School of Advanced Sciences for giving us with all the assistance and resources.

**Conflicts of Interest** All authors declared that they have no conflict of interest.

## References

[1] M. Alifano, E. Daffré, L. Brouchet, P. E. Falcoz, F. Le, P. Barthes, P. Benoit, P. A. Thomas, M. Dahan & R. Porcher (2022). Prognostic score and sex-specific nomograms to predict survival in resectable lung cancer: A French nationwide study from the Epithor cohort database. *The Lancet Regional Health - Europe*, *26*, Article ID: 100566. https://doi.org/10.1016/j.lanepe.2022.100566.

[2] F. A. Altuhaifa, K. T. Win & G. Su (2023). Predicting lung cancer survival based on clinical data using machine learning: A review. *Computers in Biology and Medicine*, *165*, Article ID: 107338. https://doi.org/10.1016/j.compbiomed.2023.107338.

[3] A. A. R. Alvarez, S. Yuming, J. Kothari, S. R. Digumarthy, N. M. Byrne, Y. Li & D. C. Christiani (2022). Sex disparities in lung cancer survival rates based on screening status. *Lung Cancer*, *171*, 115–120. https://doi.org/10.1016/j.lungcan.2022.07.015.

[4] D. Amicizia, M. F. Piazza, F. Marchini, M. Astengo, F. Grammatico, A. Battaglini, I. Schenone, C. Sticchi, R. Lavieri, B. Di Silverio, G. B. Andreoli & F. Ansaldi (2023). Systematic review of lung cancer screening: Advancements and strategies for implementation. *Healthcare*, *11*(14), Article ID: 2085. https://doi.org/10.3390/healthcare11142085.

[5] N. L. Barclay, M. Pineda Moncusí, A. M. Jödicke, D. Prieto-Alhambra, B. Raventós, D. Newby, A. Delmestri, W. Y. Man & M. Català (2024). The impact of the UK COVID-19 lockdown on the screening, diagnostics and incidence of breast, colorectal, lung and prostate cancer in the UK: A population-based cohort study. *Frontiers in Oncology*, *14*, Article ID: 1370862. https://doi.org/10.3389/fonc.2024.1370862.

[6] A. Bhattacharjee, J. Dey & P. Kumari (2022). A combined iterative sure independence screening and Cox proportional hazard model for extracting and analyzing prognostic biomarkers of adenocarcinoma lung cancer. *Healthcare Analytics*, *2*, Article ID: 100108. https://doi.org/10.1016/j.health.2022.100108.

[7] H. F. Chan, W. H. Hsu, J. P. Chen & J. H. Lee (2024). Factors associated with survival of patients with advanced lung cancer and long travel distances. *Journal of the Formosan Medical Association*, *123*(2), 273–282. https://doi.org/10.1016/j.jfma.2023.08.019.

[8] T. Chaudhry, V. Krishnan, A. E. Donaldson, Z. M. Palmisano, S. Basu, N. M. Geissen, J. M. Karush, G. C. Alex, J. A. Borgia, M. J. Liptay et al. (2023). Conditional survival analysis of patients with resected non–small cell lung cancer. *JTCVS Open*, *16*, 948–959. https://doi.org/10.1016/j.xjon.2023.09.010.

[9] M. Darshan TEJA & G. Mokesh Rayalu (2025). Exploring heart diseases survival patterns: A multifaceted approach using survival analysis techniques on time to event data. *Sigma Journal of Engineering and Natural Sciences*, *43*(4), 1307–1320. https://doi.org/10.14744/sigma.2025.00122.

[10] L. Everest, S. Blommaert, D. Tu, J. L. Pater, A. Hay, M. C. Cheung & K. K. W. Chan (2022). Validating restricted mean survival time estimates from reconstructed Kaplan-Meier data against original trial individual patient data from trials conducted by the Canadian Cancer Trials Group. *Value in Health*, *25*(7), 1157–1164. https://doi.org/10.1016/j.jval.2021.12.004.

[11] R. E. Gemine, G. R. Davies, K. Lanyon, S. E. Rees, I. Campbell & K. E. Lewis (2023). Quitting smoking improves two-year survival after a diagnosis of non-small cell lung cancer. *Lung Cancer*, *186*, Article ID: 107388. https://doi.org/10.1016/j.lungcan.2023.107388.

[12] S. Ghosh, G. Samanta & J. J. Nieto (2021). Application of non-parametric models for analyzing survival data of COVID-19 patients. *Journal of Infection and Public Health*, *14*(10), 1328–1333. https://doi.org/10.1016/j.jiph.2021.08.025.

[13] I. K. Hussein & E. A. Fadam (2024). A partial maximum likelihood method to estimate Cox model for competing risk data with application. *Journal of Economics and Administrative Sciences*, *30*(140), 450–465. https://doi.org/10.33095/619t1m17.

[14] Q. Jiang, X. Yue, H. Lei, W. Mao, Y. Li & X. Chen (2023). Comparison of survival outcomes between clinical trial participants and non-participants of patients with advanced non-small cell lung cancer: A retrospective cohort study. *Heliyon*, *9*(12), Article ID: e22660. https://doi.org/10.1016/j.heliyon.2023.e22660.

[15] S. Karim, M. S. Iqbal, N. Ahmad, M. S. Ansari, Z. Mirza, A. Merdad, S. D. Jastaniah & S. Kumar (2023). Gene expression study of breast cancer using Welch Satterthwaite $t-$test, Kaplan-Meier estimator plot and huber loss robust regression model. *Journal of King Saud University-Science*, *35*(1), Article ID: 102447. https://doi.org/10.1016/j.jksus.2022.102447.

[16] A. Lahiri, A. Maji, P. D. Potdar, N. Singh, P. Parikh, B. Bisht, A. Mukherjee & M. K. Paul (2023). Lung cancer immunotherapy: Progress, pitfalls, and promises. *Molecular Cancer*, *22*(1), Article ID: 40. https://doi.org/10.1186/s12943-023-01740-y.

[17] A. Leiter, R. R. Veluswamy & J. P. Wisnivesky (2023). The global burden of lung cancer: Current status and future trends. *Nature Reviews Clinical Oncology*, *20*(9), 624–639. https://doi.org/10.1038/s41571-023-00798-3.

[18] F. Liang, S. Zhang, Q. Wang & W. Li (2018). Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. *Annals of Oncology*, *29*(5), 1320–1324. https://doi.org/10.1093/annonc/mdy075.

[19] M. Ljunggren, C. E. Weibull, G. Palmer, E. Osterlund, B. Glimelius, A. Martling & C. Nordenvall (2023). Sex differences in metastatic surgery following diagnosis of synchronous metastatic colorectal cancer. *International Journal of Cancer*, *152*(3), 363–373. https://doi.org/10.1002/ijc.34255.

[20] M. Mustafa, A. R. JamalulAzizi, E. L. Illzam, A. Nazirah, A. M. Sharifa & S. A. Abbas (2016). Lung cancer: Risk factors, management, and prognosis. *IOSR Journal of Dental and Medical Sciences*, *15*(10), 94–101. https://doi.org/10.9790/0853-15100494101.

[21] Y. Park & Y. Chen (2023). Sample size determination and evaluation for two-stage adaptive designs of single arm clinical trials based on median event time test. *Contemporary Clinical Trials Communications*, *36*, Article ID: 101225. https://doi.org/10.1016/j.conctc.2023.101225.

[22] S. Rai, P. Mishra & U. C. Ghoshal (2021). Survival analysis: A primer for the clinician scientists. *Indian Journal of Gastroenterology*, *40*(5), 541–549. https://doi.org/10.1007/s12664-021-01232-1.

[23] N. Ratasukharom, S. A. Niwitpong & S. Niwitpong (2024). Confidence intervals for the ratio between two means of Birnbaum-Saunders distributions. *Malaysian Journal of Mathematical Sciences*, *18*(2), 301–320. https://doi.org/10.47836/mjms.18.2.06.

[24] Y. W. Sari, Gunardi, N. Y. Megawati & S. H. Hutajulu (2025). Cox-based estimation model for critical illness insurance policy for breast cancer based on the possible transition of status. *Malaysian Journal of Mathematical Sciences*, *19*(1), 35–51. https://doi.org/10.47836/mjms.19.1.03.

[25] P. Shankaraiah, G. Mokesh Rayalu & A. V. Prasad (2025). Advanced survival methods for the assessment of lung cancer survival: The influence of random and fixed effects. *Journal of Applied Mathematics & Informatics*, *43*(3), 867–883. https://doi.org/10.14317/jami.2025.867.

[26] Statista. Major cancer types by deaths worldwide 2020. Technical report Statista 2020. https://www.statista.com/statistics/288580/number-of-cancer-deaths-worldwide-by-type/.

[27] K. Sun, M. Li, M. Shang, X. Su, J. Zhao, B. Wang, C. Wu, L. Zhang, S. Yang & X. Sun (2022). Impact of genetic status on the survival outcomes of patients with clinical stage I non-small cell lung cancer with a radiological pure-solid appearance. *Lung Cancer*, *166*, 63–69. https://doi.org/10.1016/j.lungcan.2022.02.005.

[28] J. Teng, Y. Liu, J. Xia, Y. Luo, H. Zou & H. Wang (2023). Impact of time-to-treatment initiation on survival in single primary non-small cell lung Cancer: A population-based study. *Heliyon*, *9*(9), Article ID: e19750. https://doi.org/10.1016/j.heliyon.2023.e19750.

[29] B. Tharu, K. Pokhrel, G. Aryal, R. C. Kafle & N. Khanal (2021). Study of age specific lung cancer mortality trends in the US using functional data analysis. *Communications for Statistical Applications and Methods*, *28*(2), 119–134. https://doi.org/10.29220/CSAM.2021.28.2.119.

[30] X. Q. Yu, M. L. Yap, E. S. Cheng, P. J. Ngo, P. Vaneckova, D. Karikios, K. Canfell & M. F. Weber (2022). Evaluating prognostic factors for sex differences in lung cancer survival: Findings from a large Australian cohort. *Journal of Thoracic Oncology*, *17*(5), 688–699. https://doi.org/10.1016/j.jtho.2022.01.016.

[31] H. Zhu, X. You & S. Liu (2019). Multiple ant colony optimization based on pearson correlation coefficient. *IEEE Access*, *7*, 61628–61638. https://doi.org/10.1109/ACCESS.2019.2915673.