International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 1, Jan-Feb 2025

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Survival Analysis of Length of Hospitalization of Lung Cancer Patients Using the Kaplan Meier Method

M. Rizaldy Baihaqi*, Putri Masyita Qomaryah**, Ardi Kurniawan***, Elly Ana****

****(Statistics Study Program, Faculty of Science and Technology, Airlangga University, Surabaya, Indonesia

Email: ardi-k@fst.unair.ac.id)

----------**********----------

Abstract:

Lung cancer is one of the leading causes of cancer-related deaths globally, including in Indonesia, with smoking as the primary risk factor. This study aims to analyze the survival rate of lung cancer patients using the Kaplan-Meier survival analysis method. The data used in this study was obtained from the Kaggle platform, which contains secondary data on lung cancer patients. The analysis results show a significant decline in the survival probability of patients over time. The survival estimate on day one is 1.000, which decreases to 0.315 on day 23. No significant difference in survival was found between males and females, with an average survival time of around 20 days for both genders. Based on these findings, the study suggests improving lung cancer screening programs for high-risk groups, strengthening smoking prevention efforts, and conducting further research on other factors affecting patient survival. Additionally, the development of more effective treatments and the provision of psychosocial support for patients are also emphasized. This study is expected to provide insights for improving health policies, treatment, and prevention strategies for lung cancer in Indonesia.

Keywords — **Lung Cancer, Survival Analysis, Kaplan Meier**

----------**********----------

I. INTRODUCTION

Cancer is a significant global health threat, affecting countries around the world, both developing and developed. According to a 2004 report from the World Health Organization (WHO), cancer is the leading cause of death, with 7.4 million deaths and is predicted to increase to 12 million by 2030 (Christina, 2015). In Indonesia alone, lung cancer ranks third after breast and cervical cancer, with 34,783 cases (8.8%) of the total 396,914 recorded cancer cases. In addition, lung cancer is the highest cause of death, with 25,943 deaths (14.1%) out of 183,368 cancer deaths. Based on the World Cancer 2014 report by the International Agency for Research on Cancer (IARC), there were about 14 million new cancer cases in 2012, of which lung cancer accounted for about 13% of all cancer cases, followed by breast cancer at 11.9%. Lung cancer was also the highest cause of death of all cancers in the world that year.

Lung cancer is a condition in which there is an uncontrolled development of cancer cells in the lung tissue, either due to malignancies originating from outside the lung or from within the lung itself (Purba, 2015). This development is caused by mutations in genes that control the process of cell separation. These mutations can be automatic or inherited. Lung cancer often grows without showing indications until it reaches an advanced stage, and until now there is no screening procedure that can be widely used. Recommended screening for lung cancer only applies to the highest risk group of patients. People who are over 40 years old and have a history of smoking for 30 years or more, and who have stopped smoking in the last 15 years, or those over 50 years old with a history of smoking for 20 years and have at least one additional risk aspect, are included in this group.

Lung cancer is generally more common in men, especially in older age, with smoking as the main risk factor. Active smokers have a very high chance of developing lung cancer, with 80-90 percent of cases associated with smoking. Apart from smoking, other causes of lung cancer may include exposure to radiation, air pollution or toxic materials such as arsenic. The risk of lung cancer is not only experienced by male smokers, but also by women who are passive smokers (Restika, 2014). There are various factors that can cause lung cancer and lead to death. Patients diagnosed with lung cancer usually undergo treatment for a certain period, with possible outcomes of cure or death.

Some aspects that influence the survival rate of lung cancer patients include gender, age, cholesterol content, smoking history, family history, hypertension, asthma, cirrhosis, and body mass index. Research on factors affecting lung cancer survival is limited, especially in Indonesia. This research aims to identify factors that influence lung cancer survival. The information generated from this observation is in the form of patient survival duration, which is then analyzed using a statistical procedure known as survival analysis.

Survival analysis is used to examine data related to survival time until the occurrence of a certain event. In this study, the analysis was carried out to estimate the chances of survival of lung cancer patients based on data from Kaggle, using the Kaplan Meier method. The steps taken are to determine the observed data and censored data, then calculate the survival estimate S(t), estimate the failure rate H(t), then interpret (Firsawan et al., 2022). This research is expected to help develop more effective handling strategies and provide more accurate predictions.

II. RESEARCH METHODS

This study uses secondary data obtained from Kaggle.com, a platform that provides open, accessible datasets for data analysis. The dataset used in this study focuses on the survival of patients with lung cancer, taken from observational data involving several patients who have been diagnosed with lung cancer. The available data consists of two categories, namely patients with observed status (complete data) and patients with censored status, which means that their data stopped before the final event (such as death) was recorded (Smith et al., 2020).

Survival analysis was performed using the Kaplan-Meier method, which is a statistical method used to estimate the probability of patient survival within a certain period of time. This method is very effective for handling survival data involving censoring, which is data that does not record the final event because the observation is stopped at a certain time (Kleinbaum & Klein, 2017). Survival estimation using the Kaplan-Meier formula is calculated as follows:

$$\begin{aligned}\hat{S}(t) &= \hat{p}_1 \times \hat{p}_2 \times \dots \times \hat{p}_k \\ &= \prod_{j=1}^{k} \hat{p}_j \\ &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right)\end{aligned} \quad (1)$$

Where:

- $\hat{S}(t)$ is the estimated probability of survival at time t,
- n_j is the number of individuals still at risk at time t_j
- d_j is the number of individuals who experienced the event at that time.

This method is often used in medical research because it can estimate the probability of survival even if some data is censored (Bland, 2018). In addition, in this study, the hazard function is used to measure the rate of occurrence or failure at any given time, with the following formula:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \le T \le t + \delta t)}{\delta t} \quad (2)$$

This function measures the probability of an event (e.g. death) occurring in the time interval ttt, provided that the individual has survived until that time.

This study also takes into account censored data, which occurs when the timing of an event (e.g. death) is not fully observed because the observation is stopped at a certain time. There are several types of censoring that can occur in survival analysis:

1. Type I censoring: All individuals are observed at the same time, and observations are stopped at a predetermined time.
2. Type II Censorship: Censoring occurs on individuals who have experienced the smallest event in the sample.
3. Type III censoring: Individuals enter the experiment at different times.

To check the distribution of the data used, a Kolmogorov-Smirnov normality test was performed. This test compares the distribution of the observed data with the standard normal distribution. If the significance value is greater than α = 5% or 0.05, then the data can be considered normally distributed, which allows the use of further statistical methods.

III. RESULTS

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

A. Normality Assumtion Test

Before conducting the Kaplan-Meier test, it is necessary to test the normality assumption to determine whether the data is normally distributed or not.

TABLE I
NORMALITY ASSUMTION TEST

One-Sample Kolmogorov-Smirnov Test		
		Lama_Perawatan
N		261
Normal Parameters[a,b]	Mean	14,1303
	Std. Deviation	4,46937
Most Extreme Differences	Absolute	0,066
	Positive	0,066
	Negative	-0,064
Test Statistic		0,066
Asymp. Sig. (2-tailed)		.008[c]

Based on Table 1, it can be seen that the significance value is 0.008 where the value is < α (5%). So it can be concluded that the data is not normally distributed and can be tested with nonparametric analysis methods.

B. Data Analysis

After testing the normality assumption, it was found that the data was not normally distributed. So that data testing can be done using the nonparametric survival analysis method with the Kaplan-Meier method. The following table shows the results of the analysis of survival of lung cancer patients using the Kaplan-Meier method.

TABLE III
SURVIVAL ANALYSIS OF LUNG CANCER PATIENTS

i	t_i	R_i	δ_i	C_i	q_i	p_i	$S(t)$	$h(t)$
1	5	260	0	1	0,000	1,000	1,000	0,011
2	6	259	3	9	0,011	0,988	0,988	0,008
3	7	247	2	8	0,008	0,991	0,980	0,004
4	8	237	1	7	0,004	0,995	0,976	0,004
5 6	9 10	229 219	1 0	9 15	0,004 0,000	0,995 1,000	0,972 0,972	0,000 0,000
7	11	204	1	18	0,004	0,995	0,967	0,027
8	12	185	5	17	0,027	0,972	0,941	0,049
9	13	163	8	11	0,049	0,950	0,894	0,020
10	14	144	3	20	0,020	0,979	0,876	0,008
11	15	121	1	22	0,008	0,991	0,869	0,071
12	16	98	7	20	0,071	0,928	0,806	0,042
13	17	71	3	15	0,042	0,957	0,772	0,056
14	18	53	3	4	0,056	0,943	0,729	0,021
15	19	46	1	6	0,021	0,978	0,713	0,025
16	20	39	1	5	0,025	0,974	0,694	0,121
17	21	33	4	11	0,121	0,878	0,610	0,277
18	22	18	5	6	0,277	0,722	0,441	0,285
19	23	7	2	5	0,285	0,714	0,315	1,000

Keterangan Tabel:
i : observation to-*i*.
t_i : survival time of lung cancer patiens.
R_i : number of individuals alive in the i-th observation
δ_i : the number of individuals that died in the i-th observation.

C_i : the number of censored individuals in the i-th observation.

q_i : the estimated probability of individual death at the i-th observation.

p_i : the estimated probability of individual survival at the ith observation.

$S(t)$: estimation of the survival function for each individual included in the observation.

$h(t)$: estimation of the survival failure rate of each individual included in the observation.

In Table 2, it can be seen that the estimated chance of survival of lung cancer patients who survive for 5 days is 1.000. While the estimated chance of survival for lung cancer patients who survive for 23 days is 0.315. The estimated survival failure rate of lung cancer patients who survive for 5 days is 0.000, while lung cancer patients who survive for 23 days have an estimated survival failure rate of 1.000.

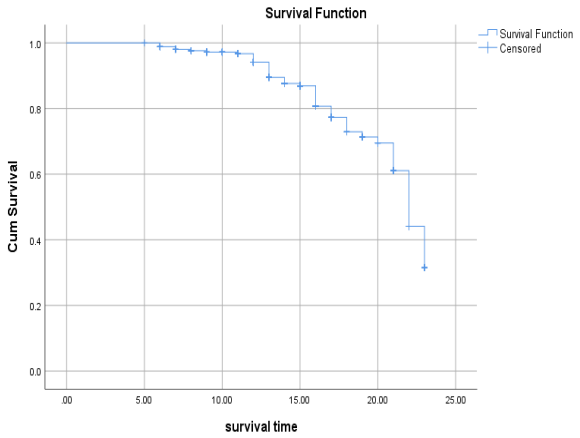

Fig. 1 Survival Plot of Lung Cancer Patients **Figures and Tables**

Figure 1 shows an overview of the characteristics of the survival plot of lung cancer patients. Based on the figure, it can be seen that the plot moves downward and the estimated chance of survival of lung cancer patients ranges from 0.315 to 1.000. Where 0.315 is the estimated chance of survival of lung cancer patients who survive for 23 days and 1,000 is the estimated chance of survival of lung cancer patients who survive for 5 days.

TABLE IIIII
SURVIVAL ANALYSIS OF LUNG CANCER P COMPARISON OF AVERAGE SURVIVAL OF LUNG CANCER PATIENTS BY GENDERATIENTS

Means for Survival Time				
Sex	Mean		95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound
Female	20,201	0,499	19,223	21,178
Male	20,214	0,460	19,312	21,115

Furthermore, in Table 3 when viewed based on gender, it can be seen that the estimated average survival of lung cancer patients with male female gender is around 20 days. While the estimated average survival of lung cancer patients with female gender is around 20 days. So the estimated survival time for women (20,201) and men (20,214) is very similar, with a very small difference between the two of only 0.013. The 95% confidence intervals for both groups also show overlapping values, indicating that there is no significant difference in the estimated survival time between women and men.

IV. DISCUSSION

The results of the Kaplan-Meier analysis showed a decrease in the estimated odds of survival of lung cancer patients as time progressed. On day 5, all observed patients were still alive (survival odds of 1,000), but by day 23, the survival odds dropped to 0.315, meaning that nearly 68.5% of patients had died before reaching day 23. This reflects that the survival of lung cancer patients decreases significantly with time.

The resulting survival plots of lung cancer patients show a consistent decrease in the chance of survival. The survival probabilities of patients ranged from 0.341 to 1.000, depending on the time of observation. These data provide a clear picture of the survival characteristics of patients, indicating that the longer the time elapses, the lower the survival chances of lung cancer patients. When compared by gender, the results showed that there was no significant difference in mean survival between males and females. The average survival for both groups was almost the same, with a difference of only 0.013 days. The 95% confidence intervals for both groups overlapped, indicating that gender did not have a

major influence on the survival duration of lung cancer patients in the sample used.

The results of this study provide useful insights in understanding the survival patterns of lung cancer patients. Although gender had no significant effect, other factors such as age, smoking habits and family history of the disease may require further analysis. This study also shows the importance of using survival analysis to project the survival of lung cancer patients, although other factors such as cancer stage and type of treatment received should also be taken into account to plan more effective treatment. However, this study has several limitations, including data that only includes patients with complete or censored data, as well as limited information regarding other factors that affect patient prognosis, such as cancer stage and type of treatment. In addition, sampling bias and limitations of secondary data also need to be considered.

V. CONCLUSIONS

Lung cancer is a significant global health problem, with a high mortality rate, especially in Indonesia. The main factors affecting the risk of lung cancer are smoking habits, both in active and passive smokers, as well as environmental factors such as air pollution and exposure to toxic materials. This study used the Kaplan-Meier method to analyze the survival of lung cancer patients and found a significant decrease in the chance of survival over time. The estimated survival of patients on day one was 1,000, which decreased to 0.315 on day 23. The results of the analysis by gender showed that there was no significant difference in survival between men and women, with a mean estimated survival of about 20 days for both genders.

Suggestions that we can provide for future research are needed to explore other factors that affect the survival of lung cancer patients, such as other medical conditions (e.g. hypertension, diabetes), nutritional status, and environmental factors.

ACKNOWLEDGMENT

The author would like to thank the Statistics Study Program, the Faculty of Science and Technology, and Airlangga University, as well as all parties involved in supporting this research and publication

REFERENCES

[1] Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., ... & Muntner, P. (2017). Heart disease and stroke statistics—2017 update: A report from the American Heart Association. Circulation, 135(10), e146– e603. https://doi.org/10.1161/CIR.0000000000000485

[2] Collett, D. (2003). Modelling survival data in medical research. Chapman & Hall.

[3] Husen, A., Suharti, C., & Hardian, H. (2016). Hubungan antara derajat nyeri dengan tingkat kualitas hidup pasien kanker paru yang menjalani kemoterapi. Jurnal Kedokteran Diponegoro (Diponegoro Medical Journal), 5(4), 545–557.

[4] Kaplan, S. J., Pelcovitz, D., & Labruna, V. (1999). Child and adolescent abuse and neglect research: A review of the past 10 years. Part I: Physical and emotional abuse and neglect. Journal of the American Academy of Child & Adolescent Psychiatry, 38(10), 1214–1222. https://doi.org/10.1097/00004583-199910000-00009.

[5] Kendal, P. C., & Hammen, C. (1998). Abnormal psychology: Understanding human problem.

[6] Kleinbaum, D. G., & Klein, M. (2005). Survival analysis: A self-learning text (2nd ed.). Springer..

[7] PDPI. (2006). Asma: Pedoman diagnosis dan penatalaksanaan di Indonesia. Perhimpunan Dokter Paru Indonesia.

[8] Price, S. A., & Wilson, L. M. (2006). Patofisiologi: Konsep klinis proses-proses penyakit. Jakarta: EGC.

[9] Purba, A. F., Wibisono, B. H., & Rachmawati, B. (2015). Pola klinis kanker paru di RSUP Dr. Kariadi Semarang periode Juli 2013–Juli 2014 (Doctoral dissertation, Faculty of Medicine).

[10] Sidney, B. (2006). The global burden of asthma: Decreasing the global burden of asthma. Chest, 130(1, Supplement), 4S–12S. https://doi.org/10.1378/chest.130.1_suppl.4S

[11] Tanto, C., Liwang, F., Hanifati, S., & Pradipta, E. A. (2014). Kapita selekta kedokteran. Jakarta: Media Aesculapius.

[12] World Health Organization (WHO). (2015). A global brief on hypertension: Silent killer, global public health crisis.

[13] World Health Organization (WHO). (2020). Air pollution and health. Retrieved from https://www.who.int/news-room/fact-sheets/detail/airpollution.

