Proceedings of CONF-CDS 2026 Symposium: Data-Centric AI Security: Securing Models, Learning Agents, and Autonomous Systems
DOI: 10.54254/2755-2721/2026.CH34579

Survival Prognostic Modeling for Lung Cancer Patients: A Comparative Analysis of Non-Parametric and Semi-Parametric Statistical Methods

Ni Li

School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu, China
42333060@smail.swufe.edu.cn

Abstract. Accurate survival prognosis is essential for personalized prognostic assessment and treatment planning in lung adenocarcinoma. This study compares non-parametric and semi-parametric statistical methods for survival prognostic modeling using clinical data from the TCGA-LUAD cohort $(n=493)$. The Kaplan-Meier estimator and the multivariate Cox proportional hazards model were applied to evaluate the prognostic roles of tumor stage, age, and gender. The dataset was divided into a training set (70%) for model fitting and a testing set (30%) for independent validation. The results show that the Kaplan-Meier estimator provides an intuitive visualization of survival differences across tumor stages, with Log-rank tests confirming significant differences among subgroups $(p<0.001)$. The Cox model identified tumor stage as the dominant independent prognostic factor. Compared with Stage I patients, Stage IV patients had a 3.58-fold higher hazard of death $(HR = 3.58, 95\%\ CI: 1.67–7.69, p<0.005)$. Although the C-index increased only slightly from 0.686 to 0.689, the Cox model offered added value through multivariate adjustment and the estimation of interpretable hazard ratios. These findings suggest that Kaplan-Meier estimation and Cox regression play complementary roles in lung cancer survival analysis.

Keywords: Survival Analysis, TCGA-LUAD, Kaplan-Meier Estimator, Cox Proportional Hazards Model, Lung Adenocarcinoma

1. Introduction

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, posing substantial challenges to clinical prognosis, treatment planning, and long-term patient management [1]. Among its major pathological subtypes, lung adenocarcinoma is highly prevalent and clinically heterogeneous, which makes accurate prognosis of survival particularly important for individualized treatment planning. The TCGA-LUAD cohort provides valuable clinical and molecular information for lung adenocarcinoma research, offering a reliable data foundation for survival prognosis modeling [2].

In biomedical statistics, survival analysis has become an essential tool for evaluating time-to-event outcomes, especially when clinical data contain censored observations. Existing studies have

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.CH34579

widely applied non-parametric methods, such as the Kaplan-Meier estimator, to describe survival probability across patient subgroups [3]. Semi-parametric methods, represented by the Cox proportional hazards model, have also been extensively used to estimate the effects of multiple prognostic factors while maintaining model interpretability [4]. However, many studies emphasize the application of a single method, whereas fewer provide a structured comparison between non-parametric and semi-parametric approaches using the same lung adenocarcinoma cohort — particularly from the perspectives of model-fitting performance, predictive accuracy, and interpretability..

To address this issue, this study compares the Kaplan-Meier estimator and the Cox proportional hazards model using clinical data from the TCGA-LUAD cohort. Specifically, the analysis focuses on survival differences across tumor stages, the prognostic effects of clinical variables, and the relative strengths and limitations of the two statistical approaches. By clarifying their applicability in censored lung cancer survival data, this study provides empirical evidence for selecting appropriate survival analysis methods and contributes to more interpretable prognostic modeling in precision oncology research.

2. Methodology and results

2.1. Data source and preprocessing

All statistical analyses and computational modeling in this study were conducted using Python and the lifelines library [5]. The study utilized retrospective clinical data from the Cancer Genome Atlas lung adenocarcinoma cohort, namely the TCGA-LUAD cohort. After preprocessing, a total of 493 eligible patients were included in the final analysis.

To ensure analytical reproducibility and data integrity, several preprocessing procedures were implemented. Patients with missing survival time or incomplete clinical staging information were excluded from the analysis. During this process, original database column names, such as OS_MONTHS and OS_STATUS, as well as fundamental clinical values, were strictly preserved to maintain data traceability. Continuous variables, such as age at diagnosis, were retained in their original scale without normalization, so that the estimated effects would remain clinically interpretable..

The primary outcome variables were overall survival time and survival status. Overall survival time was measured in months, while survival status was coded as either censored or event observed. The main covariates included age at diagnosis, gender, and tumor stage. Tumor stage was treated as a categorical clinical variable, with Stage I defined as the reference group in the Cox regression model. To evaluate predictive performance and reduce potential overfitting, the final dataset was randomly divided into a training set comprising 70% of the sample (345 patients) and a testing set comprising the remaining 30% for independent validation.

2.2. Non-parametric method: The Kaplan-Meier estimator

The Kaplan-Meier estimator was employed as the baseline non-parametric method to estimate the survival function. It is especially well-suited for censored survival data, as it requires no prior assumptions regarding the underlying distribution of survival time.. The estimated survival function is expressed as follows:

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

DOI: 10.54254/2755-2721/2026.CH34579

where t_i represents a time point at which at least one event occurs, d_i is the number of observed events at time t_i , and n_i denotes the number of individuals at risk immediately before time t_i.

In this study, the Kaplan–Meier method was primarily employed to visualize survival probability across stratified tumor stage subgroups. The Log-rank test was further applied to compare survival distributions among these subgroups [6]. The test results showed statistically significant differences in survival curves across tumor stages $(p<0.001)$, suggesting that tumor stage is strongly associated with survival outcomes in lung adenocarcinoma patients.

The strengths of the Kaplan–Meier estimator lie in its simplicity, intuitive graphical presentation, and freedom from distributional assumptions. It cannot simultaneously adjust for multiple covariates such as age, gender, and tumor stage. Therefore, although it provides an effective descriptive overview of survival differences, a semi-parametric regression model is needed for multivariate prognostic analysis.

2.3. Semi-parametric method: The Cox proportional hazards model

To evaluate the simultaneous effects of multiple prognostic factors, the Cox proportional hazards model was applied as the semi-parametric method in this study. Unlike fully parametric survival models, the Cox model does not require explicit specification of the baseline hazard function. At the same time, it allows the effects of covariates to be estimated through regression coefficients and hazard ratios. The model can be written as:

$$\lambda(t|X)=\lambda_0(t)\exp(\sum \beta_i X_i) \tag{2}$$

where $\lambda(t|X)$ denotes the hazard function at timet given covariates X, $\lambda_0(t)$ represents the baseline hazard function, and $\beta1,\beta2,\ldots,\beta p$ are regression coefficients for the included covariates. The exponential form of each coefficient, $\exp(\beta_i)$, is interpreted as the hazard ratio.

In this study, age at diagnosis, gender, and tumor stage were included in the multivariate Cox model. Tumor stage was entered as a categorical variable, with Stage I serving as the reference group. The model parameters were estimated using partial likelihood based on the training subset.

A key assumption of the Cox model is the proportional hazards assumption, which requires that the effect of each covariate on the hazard remains constant over time. This assumption was evaluated using Schoenfeld residual tests [7]. The results indicated that no statistically significant violation of the proportional hazards assumption was detected $(p>0.05)$, supporting the validity of the Cox model for subsequent hazard estimation and interpretation.

3. Results and comparative analysis

3.1. Kaplan-Meier survival curve analysis

The model fitting performance was first evaluated using Kaplan-Meier survival curves. The survival curves were stratified by tumor stage to visually inspect intergroup disparities in survival probability across patient subgroups

DOI: 10.54254/2755-2721/2026.CH34579

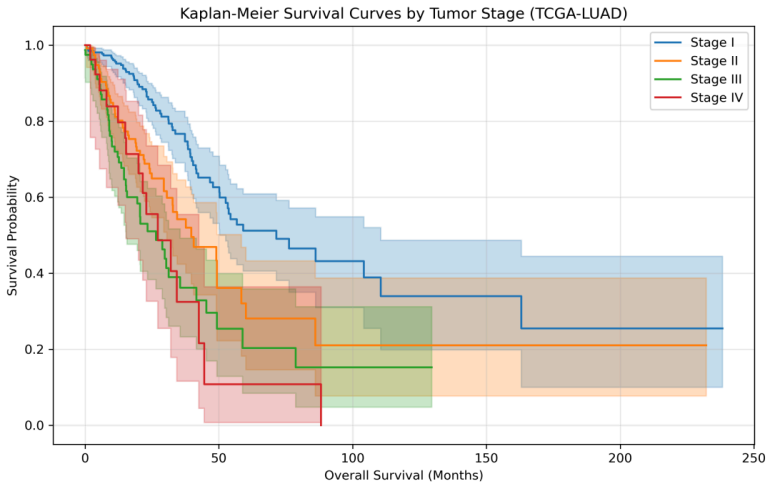

Figure 1. Kaplan-Meier survival estimates for lung cancer patients stratified by tumor stage

As shown in Figure 1, the Kaplan-Meier curves showed a clear separation among different tumor stage groups. Patients with earlier-stage disease generally maintained higher survival probabilities over time, whereas patients with advanced-stage disease experienced a more rapid decline in survival probability. This trend aligns with well-established clinical consensus that advanced tumor stage correlates with inferior prognostic outcomes.

The Log-rank test further confirmed that survival distributions differed significantly among tumor stage subgroups ($p<0.001$). These findings indicate that the Kaplan-Meier estimator is effective in presenting intuitive survival differences across clinical stages. However, as this method cannot adjust for additional covariates, the observed survival differences cannot fully capture the potential confounding effects of demographic and clinical factors.

3.2. Cox regression results

To further quantify the prognostic effects of clinical variables and control for potential confounding factors, a multivariate Cox proportional hazards model was established. As shown in Table 1, the regression results report the coefficients, hazard ratios, confidence intervals, and p-values for the included covariates.

Table 1. Multivariate Cox proportional hazards regression model

Covariate	Coefficient (β)	Hazard Ratio (HR)	95% CI for HR	p-value
Age at Diagnosis	0.01	1.01	[0.99, 1.03]	0.41
Gender (Male=1)	0.28	1.33	[0.93, 1.90]	0.12
		Tumor Stage*		
Stage II vs I	0.58	1.79	[1.15, 2.79]	0.01
Stage III vs I	1.05	2.86	[1.83, 4.46]	< 0.005
Stage IV vs I	1.28	3.58	[1.67, 7.69]	< 0.005

Note: Stage I was used as the reference group. HR = Hazard Ratio; CI = Confidence Interval.

DOI: 10.54254/2755-2721/2026.CH34579

The Cox regression results identified tumor stage as the dominant independent prognostic factor in this cohort. Compared with Stage I patients, the hazard ratios increased progressively with tumor stage. Specifically, Stage II patients had a 1.79-fold higher hazard of death than Stage I patients, while Stage III patients had a 2.86-fold higher hazard. Stage IV patients showed the highest risk, with a hazard ratio of 3.58. This hierarchical elevation in hazard ratios signifies that advanced tumor stage is strongly associated with worse survival outcomes..

In contrast, age at diagnosis and gender did not show statistically significant associations with survival outcomes in this model. Age had an HR of 1.01 with a p-value of 0.41, while male gender had an HR of 1.33 with a p-value of 0.12. Although these variables were controlled in the model, their effects were not statistically significant at the conventional 0.05 level. This result further indicates that tumor stage played the primary role in explaining survival differences within the TCGA-LUAD cohort.

It should also be noted that the confidence interval for Stage IV was relatively wider than those for Stage II and Stage III. This may reflect the limited sample size or greater heterogeneity within the Stage IV subgroup. Therefore, while the estimated effect of Stage IV is clinically meaningful and statistically significant, its uncertainty should still be considered when interpreting the result.

3.3. Predictive accuracy and model validation

The predictive performance of the models was evaluated using Harrell's Concordance Index, or C-index, a widely used metric for assessing the discriminative ability of prognostic survival models [8]. The C-index quantifies the ability of a survival model to accurately rank patients based on their survival risk. A higher C-index indicates better discriminative ability.

A baseline discriminative performance was obtained using tumor stage as a single stratification factor, resulting in a C-index of 0.686. In comparison, the multivariate Cox proportional hazards model achieved a C-index of 0.689 after incorporating tumor stage, age, and gender. Although the numerical gain remained marginal, the Cox model offered substantial methodological merits by incorporating multiple clinical covariates and deriving interpretable hazard ratios..

This finding suggests that the advantage of the Cox model in this study does not mainly lie in a large improvement in predictive accuracy. Instead, its main contribution lies in multivariate adjustment and risk interpretation. The model allows researchers to estimate the independent effect of each variable while controlling for others, which cannot be achieved by the Kaplan-Meier estimator. Accordingly, the Cox model provides a more comprehensive framework for survival prognostic analysis, particularly when the research aim is to identify and quantify multiple prognostic factors.

3.4. Methodological comparison between Kaplan-Meier and Cox models

As shown in Table 2, the Kaplan-Meier estimator and the Cox proportional hazards model differ in model type, main purpose, covariate adjustment, assumptions, predictive evaluation, strengths, and limitations.

DOI: 10.54254/2755-2721/2026.CH34579

Table 2. Comparison between Kaplan-Meier Estimator and Cox proportional hazards model

Aspect	Kaplan-Meier Estimator	Cox Proportional Hazards Model
Model type	Non-parametric method	Semi-parametric regression model
Main purpose	Descriptive survival estimation	Multivariate prognostic modeling
Main output	Survival curve and survival probability	Hazard ratio and covariate effect
Distribution assumption	Not required	Baseline hazard unspecified
Covariate adjustment	Limited to stratified comparison	Allows simultaneous adjustment
Model assumption	Independent censoring	Proportional hazards assumption
Predictive evaluation	C-index = 0.686 using stage stratification	C-index = 0.689 using multivariate covariates
Strength	Intuitive and easy to interpret visually	Provides adjusted and interpretable risk estimates
Limitation	Cannot handle multiple covariates simultaneously	Requires PH assumption and careful validation

The comparison shows that the Kaplan-Meier estimator is more suitable for preliminary and descriptive survival analysis, especially when the goal is to visualize survival differences among groups. It features simplicity and methodological transparency, and is highly effective for illustrating overall patient survival patterns stratified by tumor stage.

However, the Cox proportional hazards model is more appropriate when the research objective is to evaluate multiple prognostic variables at the same time. Although its improvement in C-index was small in this study, it provided additional statistical interpretability by estimating adjusted hazard ratios. This renders the Cox model more informative for quantifying the relative contributions of clinical risk factors.

Overall, the results suggest that the two methods should not be viewed as mutually exclusive. Instead, they serve complementary roles in survival analysis. The Kaplan-Meier estimator provides a clear visual foundation, while the Cox model extends the analysis by incorporating multivariate adjustment and quantitative risk estimation. For survival prognosis of lung adenocarcinoma, the integrated application of the two methods yields a more comprehensive and interpretable analytical framework.

4. Conclusion

This study conducted a comparative analysis of non-parametric and semi-parametric statistical methods for modeling the survival prognosis of lung adenocarcinoma patients using clinical data from the TCGA-LUAD cohort. By applying the Kaplan-Meier estimator and the Cox proportional hazards model, this research evaluated the prognostic roles of tumor stage, age, and gender. The results show that the Kaplan-Meier method provides an intuitive and distribution-free visualization of survival probability across tumor stages, making it suitable for preliminary descriptive survival analysis. However, its lack of capacity to adjust for multiple covariates restricts its applicability in complex multivariate prognostic modeling.

The Cox proportional hazards model identified tumor stage as the dominant independent prognostic factor in this cohort. Compared with Stage I patients, Stage IV patients showed a 3.58-fold higher hazard of death, indicating a clear hierarchical risk structure across clinical stages. Although the improvement in predictive accuracy was modest, with the C-index increasing from

DOI: 10.54254/2755-2721/2026.CH34579

0.686 to 0.689, the Cox model demonstrated additional methodological value through multivariate adjustment and interpretable hazard ratio estimation. Accordingly, the strengths of this semi-parametric approach lie not only in discriminative capability, but also in its capacity to quantify the relative impacts of clinical risk factors under a rigorous statistical framework.

Overall, the findings suggest that Kaplan-Meier estimation and Cox regression play complementary roles in lung cancer survival analysis. The former is useful for visualizing survival differences, while the latter is more appropriate for evaluating multiple prognostic factors simultaneously. Nevertheless, this study has several limitations. The analysis was based on a single TCGA-LUAD cohort, which may restrict generalizability, and only a limited number of clinical variables were included. Future research may integrate genomic profiles, treatment-related data, and advanced survival models including LASSO-Cox regression and Random Survival Forests to enhance personalized prognostic performance.

References

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

[2] The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511(7511), 543–550. doi: 10.1038/nature13385

[3] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. doi: 10.1080/01621459.1958.10501452

[4] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x

[5] Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317. doi: 10.21105/joss.01317

[6] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50(3), 163–170.

[7] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69(1), 239–241. doi: 10.1093/biomet/69.1.239

[8] Harrell, F. E., Jr., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4), 361–387. doi: 10.1002/(SICI)1097-0258(19960229)15: 4< 361: : AID-SIM168> 3.0.CO; 2-4