Integrated analysis of therapeutic strategies and prognostic factors in advanced lung adenocarcinoma: Retrospective study with emphasis on gene assays, multimodality treatment approaches and predictive machine learning models

SHINGCHERN YOU[1], SHIUAN-WEN CHEN[2], IRENE CHEN[3] and WEI-TEING CHEN[4,5]

[1]Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106, Taiwan, R.O.C.; [2]Department of Electrical and Computer Engineering, Faculty of Applied Science and Engineering, University of Toronto, Toronto, ON M5S 2E4, Canada; [3]Faculty of Applied Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [4]Division of Chest Medicine, Department of Medicine, Cheng-Hsin General Hospital, Taipei 112, Taiwan, R.O.C.; [5]School of Medicine, National Defense Medical Center, Taipei 114, Taiwan, R.O.C.

Received June 11, 2025; Accepted January 22, 2026

DOI: 10.3892/ol.2026.15514

Abstract. Patients with advanced lung adenocarcinoma have a range of treatment options, including targeted therapy and gene assay-guided chemotherapy. The aim of the present study was to investigate the prognostic factors and treatment-related variables influencing overall survival (OS), and to develop and validate machine learning models to predict OS in patients with advanced lung adenocarcinoma. Data on the clinical features and treatment strategies of patients with advanced lung adenocarcinoma were collected, and the effects of these variables on OS were analyzed. A total of 575 patients were included in a retrospective analysis. Among these patients, 38.4% had bone metastases, 17.9% had liver metastases, 36.0% had brain metastases and 28.9% had lung metastases. Smoking status, wild-type EGFR and older age were identified as poor prognostic factors. Patients with EGFR mutations demonstrated a prolonged OS. However, differences in OS were not observed among patient subgroups stratified by programmed death-ligand 1 expression and by common EGFR mutation subtypes, namely exon 19 deletion and L858R. First-line treatment with the tyrosine kinase inhibitor afatinib was associated with improved OS compared with that of patients treated with erotinib or gefitinib. In addition, combination therapy with the angiogenesis inhibitor bevacizumab had a positive impact on OS. Finally, several machine learning models were validated to predict OS using clinical and molecular features, and their performance was assessed using the concordance index. These findings highlight the importance of molecular profiling and individualized treatment strategies in optimizing OS for patients with advanced lung adenocarcinoma. Furthermore, the validated machine learning models may serve as useful tools for risk stratification and personalized prognostic assessment to support clinical decision-making.

Correspondence to: Dr Wei-Teing Chen, Division of Chest Medicine, Department of Medicine, Cheng-Hsin General Hospital, 45 Cheng-Hsin Street, Taipei 112, Taiwan, R.O.C.
E-mail: chenweiteing@gmail.com

Key words: lung adenocarcinoma, tyrosine kinase inhibitor, overall survival, machine learning, concordance index

Introduction

Lung cancer is a leading cause of cancer-related mortality worldwide and is a heterogeneous disease, requiring personalized therapeutic approaches to improve patient outcomes. As the predominant histological subtype of non-small cell lung cancer (NSCLC), advanced lung adenocarcinoma presents a considerable clinical challenge, with the 5-year survival rate of patients with stage IIIB/IV disease remaining <20% (1). Consequently, advanced adenocarcinoma has become a major focus of precision oncology, with treatment decisions increasingly guided by genetic mutations, such as epidermal growth factor receptor (EGFR) mutations, anaplastic lymphoma kinase (ALK) gene rearrangements and/or programmed death-ligand 1 (PD-L1) expression (1-3). Novel therapeutic strategies involving tyrosine kinase inhibitors (TKIs) and immune checkpoint inhibitors have been associated with improved overall survival (OS) in randomized clinical trials, when compared with conventional chemotherapy (2,4).

The present study integrates a decade of clinical data (2010-2019), with machine learning methodologies to identify key prognostic factors and establish predictive models. This has several notable innovations: First, it combines traditional statistical analyses with five machine learning models, including decision trees, random forests and gradient boosting, to conduct comparisons of survival prediction performance. Second, the study includes 26 clinical characteristic variables encompassing molecular biomarkers,

treatment modalities and metastatic sites. Third, it uses a concordance index (C-index) dynamic assessment framework, to overcome the limitations associated with traditional static performance indicators.

The present study examined the prognostic relevance of specific genetic alterations, including EGFR mutations and PD-L1 expression, to OS. Mutant EGFR was a particular focus, with a nuanced exploration of different EGFR mutation subtypes, including exon 19 deletions and L858R mutations, to further characterize the molecular underpinnings of lung adenocarcinoma.

Materials and methods

Study design and setting. Data of patients with lung cancer were collected from Cheng-Hsin General Hospital (Taipei, Taiwan) between January 2010 and December 2019. The present retrospective study of the data was approved by the Institutional Review Board of Cheng-Hsin General Hospital [approval no. CHGH-IRB (922)111-011; date of approval, January 27, 2022], and the requirement for informed consent was waived by the ethics committee. Clinical data were collected from a total of 1,875 patients with lung cancer, of whom 1,190 were diagnosed with adenocarcinoma. Among these patients with adenocarcinoma, 575 subjects (male/female, 291/284; mean age, 68.94±12.18 years) were initially diagnosed with advanced stage [Lung American Joint Committee on Cancer (AJCC) 8th edition staging system; stage IIIB or IV] disease (5). Due to not having undergone surgical tumor excision, patients with advanced lung adenocarcinoma may vary in prognosis according to their tumor genetic characteristics and the therapeutic strategies chosen by clinicians. In addition to the disease stage and histological subtype described, patients were included if they had a confirmed diagnosis of lung adenocarcinoma and available clinical data for analysis. No other specific inclusion or exclusion criteria were applied.

Smoking status was categorized as never smoker, current smoker, former smoker (quit <10 years ago) and long-term former smoker (quit >10 years ago). Compound mutations were defined as the presence of at least two gene mutations. Various characteristics of the patients with advanced adenocarcinoma were analyzed using multivariate regression analysis, including the tumor-node-metastasis stage based on the Lung AJCC 8th edition staging system, the location of the primary tumor, smoking status, smoking quantity (pack-year) and biomarkers, including EGFR mutation status and subtype, ALK status, PD-L1 expression and CEA levels. EGFR mutation status and subtype were determined using quantitative PCR or next-generation sequencing performed on tumor tissue specimens as part of routine clinical care. ALK rearrangement status was assessed by immunohistochemistry. PD-L1 expression was evaluated by immunohistochemistry using the anti-PD-L1 monoclonal antibody clone 22C3 (PD-L1 IHC 22C3 pharmDx; Agilent Technologies, Inc.) and reported as the tumor proportion score. Serum CEA levels were measured using standard clinical immunoassays. All biomarker data were obtained retrospectively from patients' medical records and were not generated specifically for the present study. The artificial intelligence-based algorithms were applied to structured clinical data to evaluate OS and to explore potential associations with treatment strategies.

Machine learning models. Five models were compared: i) A simple rule-based model comprising a decision tree incorporating EGFR variant patterns; ii) a random survival forest (RSF); iii) a Cox proportional hazards model; iv) a support vector machine (SVM)-based survival model; and v) a gradient-boosted survival (GBS) analysis. Model performance was evaluated using the C-index, where a C-index of <0.5 indicates random prediction, and a value >0.7 indicates good concordance between the predicted and observed survival times. For each model, the mean and standard deviation of the C-index were calculated across repeated runs to assess the average predictive accuracy and the robustness of the model under repeated random data splits. In each run, the dataset was randomly split into a training set (70%) and testing set (30%). Model performance was also evaluated using the time-dependent area under the curve (AUC) of the receiver operating characteristic curve.

Simple rule-based model. A simple rule-based decision tree model was implemented to establish a lower-bound performance benchmark. EGFR status was encoded as a categorical variable in the dataset (0, unknown; 1, wild-type; 2, mutant EGFR) according to the predefined coding scheme. The model assigned a fixed risk score based solely on the EGFR feature: Patients with EGFR=2 were assigned a predicted risk score of 0.4 (lower risk), and all other patients were assigned a risk score of 0.6 (higher risk). Because the C-index depends only on relative risk ranking and is invariant to monotonic transformations, the specific numeric values are arbitrary and serve only to distinguish lower-risk (0.4) vs. higher-risk (0.6) groups. Model performance was evaluated using the C-index to assess whether this heuristic approach captured any prognostic signal in the data.

RSF. RSFs extend the random forest framework to censored survival data by growing an ensemble of survival trees using bootstrap samples. Each tree recursively partitions the feature space to maximize survival differences using log-rank statistics. Ensemble predictions are obtained by aggregating survival probabilities across all trees. In the present study, the ExtraSurvivalTrees package (scikit-survival; version 0.26.0; https://scikit-survival.readthedocs.io/en/stable/index.html) in Python (version 3.9.7; https://www.anaconda.com/products/distribution) was implemented with 100 trees (n_estimators=100). RSFs are non-parametric models capable of capturing nonlinear relationships and feature interactions without requiring the proportional hazards assumption.

Cox proportional hazards model. The Cox proportional hazards model is a semi-parametric regression approach that relates covariates to the hazard function. The CoxnetSurvivalAnalysis package (scikit-survival 0.26.0 in the Python library) was used, which extends the classical Cox model by incorporating L1 and L2 regularization through an elastic net penalty. This regularization enables variable

selection and stabilizes coefficient estimates, particularly in high-dimensional feature spaces.

SVM. SVMs formulates survival prediction as a ranking problem, aiming to learn a function f(x) that preserves the ordering of survival times. The FastKernelSurvivalSVM package (implemented using the scikit-survival Python library; version 0.26.0) with a linear kernel was employed. This approach allows the modeling of complex relationships while maintaining computational efficiency.

GBS. Gradient boosting for survival data combines multiple weak learners, typically decision trees, in a stage-wise manner to optimize a loss function derived from the Cox partial likelihood. The GradientBoostingSurvivalAnalysis package with 100 base learners was used (sksurv.ensemble. GradientBoostingSurvivalAnalysis was implemented using the scikit-survival in the Python library). This approach sequentially refines the prediction by emphasizing samples with larger prediction errors, resulting in a strong performance in heterogeneous and high-dimensional survival datasets.

Statistical analysis. Categorical variables are presented as frequency and percentage and continuous variables are presented as mean ± standard deviation. Survival outcomes were analyzed using time-to-event methods. Kaplan-Meier curves were generated to visualize differences in survival patterns between and among patient subgroups. The survival distributions of two or more independent groups were compared using the log-rank test. However, as the proportional hazards assumption of the log-rank test may be violated when survival curves cross, the two-stage hazard rate comparison (TSHRC) method was used when non-proportional hazards were suspected, including situations with late-stage curve crossover. In this method, a log-rank test was performed at stage I to assess OS differences (α=0.05), and if the stage I test did not indicate statistical significance, a stage II maximum-type test sensitive to late or transient crossing hazards was then performed. Finally, an overall two-stage P-value was calculated. A stabilizing constant (ε=0.1) was applied to ensure numerical stability in hazard estimation. These analyses were performed using the TSHRC package in R (version 4.5.2; https://cran.r-project.org/web/packages/TSHRC).

The CoxPHSurvivalAnalysis model implemented in the Python scikit-survival library was used to assess the effects of clinical variables on OS and to generate risk predictions for survival outcomes. Associations between survival and clinical variables, including sex, age, smoking status, EGFR mutation status, CEA level and tumor location, were assessed using multivariable Cox proportional hazards regression to appropriately account for censored time-to-event data. The strength and sources of collinearity among variables in the multivariable Cox model were assessed using the Belsley-Kuh-Welsch collinearity diagnostics. Heteroscedasticity and deviation from normality were detected using the White test and Shapiro-Wilk test, respectively. Patients with missing data were excluded from the multivariable Cox analysis. Statistical analyses were performed using EasyMedStat (version 3.17; www.easymedstat.com). P<0.05 was considered to indicate a statistically significant result.

Results

Characteristics and OS of patients with advanced adenocarcinoma of the lung. A total of 575 patients with advanced lung adenocarcinoma were included in the present study. The median age of the patients was 70 years. Patient characteristics are shown in Table I. Bone metastasis was detected in 38.4% of patients, liver metastasis in 17.9%, brain metastasis in 36% and contralateral lung metastasis in 28.9%.

A total of 26 characteristic variables were included in the analysis. A Cox proportional hazards model was fitted separately for each variable, and its predictive power for OS was recorded during training. The predictive power was determined using this metric and presented in Table II, with predictive power <0.5 indicating a poor predictive performance. The five patient characteristics with the highest predictive power were, in descending order, EGFR mutation status, EGFR variant type, second-line treatment regimen, age and first-line treatment regimen.

Multivariable analysis was performed using the EasyMedStat online application to estimate and compare the effect coefficients of different variables on the survival of patients with advanced lung adenocarcinoma (Table III). Older age at diagnosis was associated with poorer survival, suggesting that the cancer had a higher malignancy and led to a poorer prognosis in older patients. Patients with left-sided lung cancer tended to have a worse prognosis than those with right-sided tumors, although this finding was not statistically significant. Smoking history was associated with adverse survival outcomes, with the poorest OS observed among those patients with former smoker status.

Effect of EGFR mutations on OS. As shown in Fig. 1A, patients with EGFR mutations had a longer OS than those without EGFR mutations (n=229 vs. 218; median OS, 24.7 vs. 11 months, respectively). Patients with an unknown EGFR mutation status had a median OS of only 7.03 months (n=118). The lack of knowledge of mutation status was primarily attributable to insufficient tumor tissue being available for molecular testing, and the patients refusing to undergo another biopsy.

In the analysis of the impact of EGFR mutation subtypes on OS, no statistically significant difference in OS was detected between the patients with the common EGFR mutation subtypes exon 19 deletion and L858R (n=104 vs. 100; median OS, 26.4 vs. 23.1 months, respectively). Patients with other less common EGFR variants had a longer median OS time (n=26; median OS, 39.3 months), as shown in Fig. 1B.

Fig. 1C shows the effect of four uncommon EGFR mutation subtypes on OS, including exon 18 gene mutation, exon 20 insertion, L861Q and compound mutations (n=6 vs. 4 vs. 3 vs. 13; median OS, 8.6, 5.8, 50.9 and 67.9 months, respectively). Among these subgroups, patients with exon 20 insertions exhibited the shortest survival.

Effect of ALK positivity on OS. A total of 255 patient samples were tested for ALK rearrangements, of which 10 were positive, accounting for 3.92% of all tested patients. However, the median OS was not reached, as shown in Fig. 1D.

Effect of therapeutic strategies on OS. Fig. 2A shows the impact of different first-line targeted therapies on the OS of

Table I. Demographic data of patients with advanced lung adenocarcinoma.

Variables	N (%)	Mean ± SD	Median	Min-max	[95% CI]
Sex					
Male	291 (50.6)	-	-	-	-
Female	284 (49.4)	-	-	-	-
Age, years	575	68.94±12.18	70.00	30.00-96.00	[67.940 to 69.931]
Height, cm	477	160.17±8.91	160.00	131.00-187.00	[159.374 to 160.974]
Weight, kg	480	59.39±12.42	59.00	27.00-113.00	[58.275 to 60.496]
Smoking quantity, pack-year	441	12.60±24.26		0.00-160.00	[10.338 to 14.866]
CEA, ng/ml	565	481.16±2,735.94	25.70	1.00-42,412.00	[255.558 to 706.758]
EGFR status					
Mutated	229 (39.8)	-	-	-	-
Wild type	218 (37.9)	-	-	-	-
Unknown	118 (20.6)	-	-	-	-
ALK positive	10 (1.7)	-	-	-	-
Metastases					
Bone	221 (38.4)	-	-	-	-
Liver	103 (17.9)	-	-	-	-
Brain	207 (36.0)	-	-	-	-
Lung	166 (28.9)	-	-	-	-
Overall survival, months	575	19.47±21.50	11.43	0.07-112.17	[17.710 to 21.225]

EGFR, epidermal growth factor receptor; ALK, anaplastic lymphoma kinase.

patients with EGFR mutations. First-line treatment with the second-generation EGFR TKI afatinib was associated with a longer OS than that obtained with the first-generation EGFR TKIs gefitinib and erlotinib (n=45 vs. 97 and 104; median OS, 33.4 vs. 19.9 and 21.3 months, respectively; P=0.0358).

In the treatment of clinically advanced lung adenocarcinoma, angiogenesis inhibitors have been shown to improve progression-free survival (PFS) (6). In the present study, the OS of patients who received three or more courses of bevacizumab was longer than that of patients who had not been treated with bevacizumab or had received fewer than three cycles (n=71 vs. 504; median OS, 23.9 vs. 13.4 months; respectively; hazard ratio, 0.74; CI, 0.57 to 0.96), as shown in Fig. 2B. The TSHRC analysis revealed stage-I P=0.039 and final two-stage P=0.0253, confirming that the survival advantage was statistically significant after adjustment for potential hazard crossing.

Regarding the effect of angiogenesis inhibitors, the present data suggested that the use of bevacizumab in late treatment lines, defined as the third line or later, was associated with a longer OS than use in earlier lines; (second line or earlier vs. third line or later; n=35 vs. 36; median OS, 16.4 vs. 28.1 months; hazard ratio, 1.7; 95% CI, 1.0 to 2.9), as shown in Fig. 2C. However, this apparent advantage was not statistically significant when hazard crossing was accounted for, as reflected by stage I log-rank P=0.138, stage II P=0.056 and final two-stage P=0.080.

PD-L1 expression in lung adenocarcinoma tissue was assessed to evaluate the potential benefit of immunotherapy. The association between PD-L1 expression and OS was analyzed (Fig. 3). Immunohistochemistry using a 22C3 PD-L1 monoclonal antibody was performed in 137 patients. Of these patients, 63 had <1% PD-L1, 56 had 1-50% PD-L1 and 18 had >50% PD-L1 staining. The median OS times of these groups were 39.7, 34.1 and 84.5 months respectively; however, the differences among the groups were not statistically significant.

Machine learning models for survival analysis. In survival prediction, patients with a higher estimated risk are expected to have shorter survival times. Model performance was evaluated using the C-index, which predicts the proportion of concordant patient pairs. Table IV summarizes the average C-index values for the different machine learning models. Among the evaluated models, the SVM achieved the highest C-index (~0.7018). Evaluation time points were defined at 3-month intervals from 1 to 60 months, providing a uniform temporal grid for performance assessment throughout the follow-up period. For each subject in the test set, the RSF, GBS and Cox proportional hazard models were trained separately to estimate individual cumulative hazard functions (Fig. 4A-C). The GBS model maintained a relatively stable AUC ranging from 0.65 to 0.75. By contrast, the AUC values of the RSF and Cox models gradually increased over 40 months, and reached their highest values at ~60 months. To further illustrate the individualized survival predictions generated by the RSF model, survival functions were estimated for a subset of test patients. Specifically, the first five subjects from the test set were selected, and their Kaplan-Meier survival curves are shown in Fig. 4D. The figure demonstrates distinct survival patterns among these subjects, with clear differences in survival probability over the entire follow-up period, highlighting heterogeneity in individual survival outcomes.

Table II. Predictive power of 26 characteristics for overall survival in patients with advanced lung adenocarcinoma.

Characteristics	Predictive power
EGFR mutation	0.623
EGFR variant type	0.607
Second-line treatment agent	0.589
Age	0.580
First-line treatment agent	0.573
Third-line treatment agent	0.570
Smoking quantity	0.564
PD-L1	0.557
Metastasis	0.548
ALK	0.548
CEA	0.542
Betel nut chewing	0.538
Weight	0.533
Radiotherapy	0.532
Lymph node involvement	0.532
Tumor size	0.528
Brain metastasis	0.526
Bone metastasis	0.524
Staging	0.522
Sex	0.521
Height	0.515
Liver metastasis	0.510
Smoking status	0.507
Primary tumor site	0.507
Lung metastasis	0.490
Alcohol consumption	0.484

EGFR, epidermal growth factor receptor; PD-L1, programmed death-ligand 1; ALK, anaplastic lymphoma kinase.

Discussion

The comprehensive retrospective analysis of 575 patients in the present study provides a nuanced understanding of the complex treatment landscape and prognostic factors associated with advanced lung adenocarcinoma. Numerous previous studies have used PFS as a surrogate endpoint because results for this can be obtained relatively quickly; however, there is inevitably a temporal bias in PFS due to its intermittent assessment (7). By contrast, the present study used OS as the evaluation outcome, as the time of death is definitive and accurately recorded, thereby providing a more accurate reflection of clinical outcomes (8).

One notable aspect of the present study is the diversity of metastatic sites included within the patient cohort. The identified prognostic factors, including smoking status, EGFR mutation status and age, contribute to a deeper understanding of patient outcomes. The recognition that certain demographic and molecular characteristics are poor prognostic indicators underscores the importance of personalized and targeted interventions. For example, the presence of bone metastases was identified as a factor associated with poor outcomes, and possibly poor clinical condition, indicating that heightened clinical attention and tailored therapeutic approaches are warranted for this subgroup.

According to previous studies, approximately one-third of patients with stage IV lung adenocarcinoma have bone metastasis, and nearly one-quarter have lung metastasis, with liver metastasis being associated with the poorest prognosis, followed by bone metastasis (9,10). In the present study, the OS of patients with advanced lung adenocarcinoma was comparable between men and women; although women exhibited a slightly worse prognosis, the difference between sexes was not statistically significant. The most common metastatic locations in the present cohort were the bones, brain, lungs and liver. The prognostic impact was indicated to be in the order of brain, bone, liver and lung. The discrepancy between these findings and those of previous studies may be attributable to differences in the incidence of metastasis at each location, which may affect the predictive ability for survival outcomes.

It has previously been reported that smoking volume is associated with reduced OS in patients with advanced lung adenocarcinoma (11). Consistent with this, the present study showed that the patients who had quit smoking within the last 10 years and current smokers had a worse OS compared with that of patients who had never smoked. Whether smoking contributes to the development of drug resistance requires further investigation. In addition, left-sided primary lung cancer showed a non-significant trend toward being associated with poorer OS, possibly due to its proximity to critical structures such as the heart, aorta and esophagus.

Among all the patient characteristics evaluated, EGFR mutation status was found to have the highest predictive power for OS. A molecular epidemiological analysis of lung adenocarcinoma conducted in Taiwan in 2016, based on 1 year of cumulative EGFR gene testing, revealed that 56.1% of patients were aged ≤45 years and 60.6% harbored EGFR mutations (12). By contrast, the present study focused primarily on patients with advanced lung adenocarcinoma, of which ~40% had confirmed EGFR gene mutations, while 20.6% had unknown or untested EGFR status. The therapeutic efficacy of targeted agents for EGFR-mutated lung adenocarcinoma is known to vary according to mutation subtype, with common subtypes including exon 19 deletion and L858R; first-line targeted therapy has been shown to result in a longer PFS in cases with exon 19 deletions (13). In the present study, the population proportions and OS outcomes of these two mutation subtypes were comparable, with no statistically significant difference observed. The cohort also included a small group of patients with unknown EGFR mutation status. Patients with unknown EGFR mutation status were analyzed as a separate category and had the shortest median OS, which likely reflected their poorer clinical condition and inability to undergo repeat molecular testing or aggressive treatment at diagnosis.

Among rare EGFR mutation subtypes, exon 20 insertions were observed in a small number of cases in the present study; they were found to be associated with a short OS (median, 5.8 months) and are reportedly associated with a poor response to targeted drugs and chemotherapy (14). The second most common rare mutation was exon 18 mutation G719X, which was also associated with a short survival

Table III. Multivariate regression analysis of variables associated with survival in patients with advanced lung adenocarcinoma.

Variables	Regression coefficient, β [95% CI]	P-value
Intercept	35.04 [24.1 to 45.97]	<0.0001
Age, per 1-unit increase	-0.252 [-0.396 to -0.107]	6.82×10^{-4}
Tumor location		
Right	Reference	
Left	-3.26 [-6.77 to 0.24]	0.0678
Smoking status		
Never	Reference	
Current	-3.62 [-9.24 to 1.99]	0.205
Long-term former	1.52 [-5.87 to 8.91]	0.686
Former	-6.43 [-12.28 to -0.58]	0.0313
Unknown	-8.36 [-13.83 to -2.9]	0.00278
EGFR		
Wild type	Reference	
Mutation	9.23 [5.17 to 13.29]	<0.0001
Unknown	-0.961 [-5.59 to 3.67]	0.683
CEA, per 1-unit increase	0.0114 [-6.81×10^{-4} to 0.0236]	0.0643
Sex		
Female	Reference	
Male	1.96 [-2.1 to 6.01]	0.343
Smoking quantity, per 1-unit increase	-0.181 [-0.243 to -0.119]	<0.0001

EGFR, epidermal growth factor receptor.

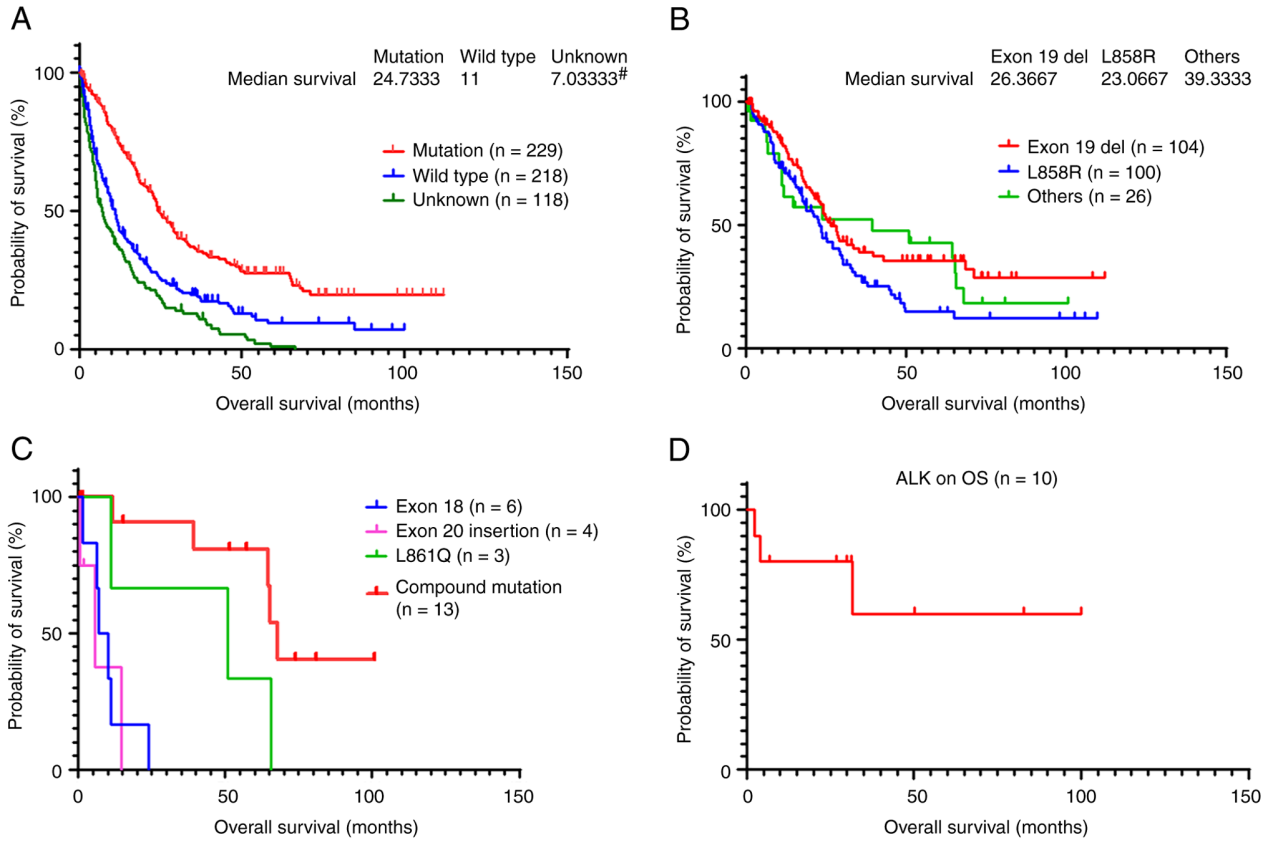

Figure 1. OS of patients with advanced lung adenocarcinoma according to EGRF and ALK status. Kaplan-Meier plots for OS by (A) EGFR mutation status, (B) common EGFR mutation subtypes, (C) uncommon EGFR mutations and (D) ALK-positive patients. The median OS for ALK-positive patients was not reached. #P<0.001 among the groups, OS, overall survival; EGFR, epidermal growth factor receptor; ALK, anaplastic lymphoma kinase; del, deletion.

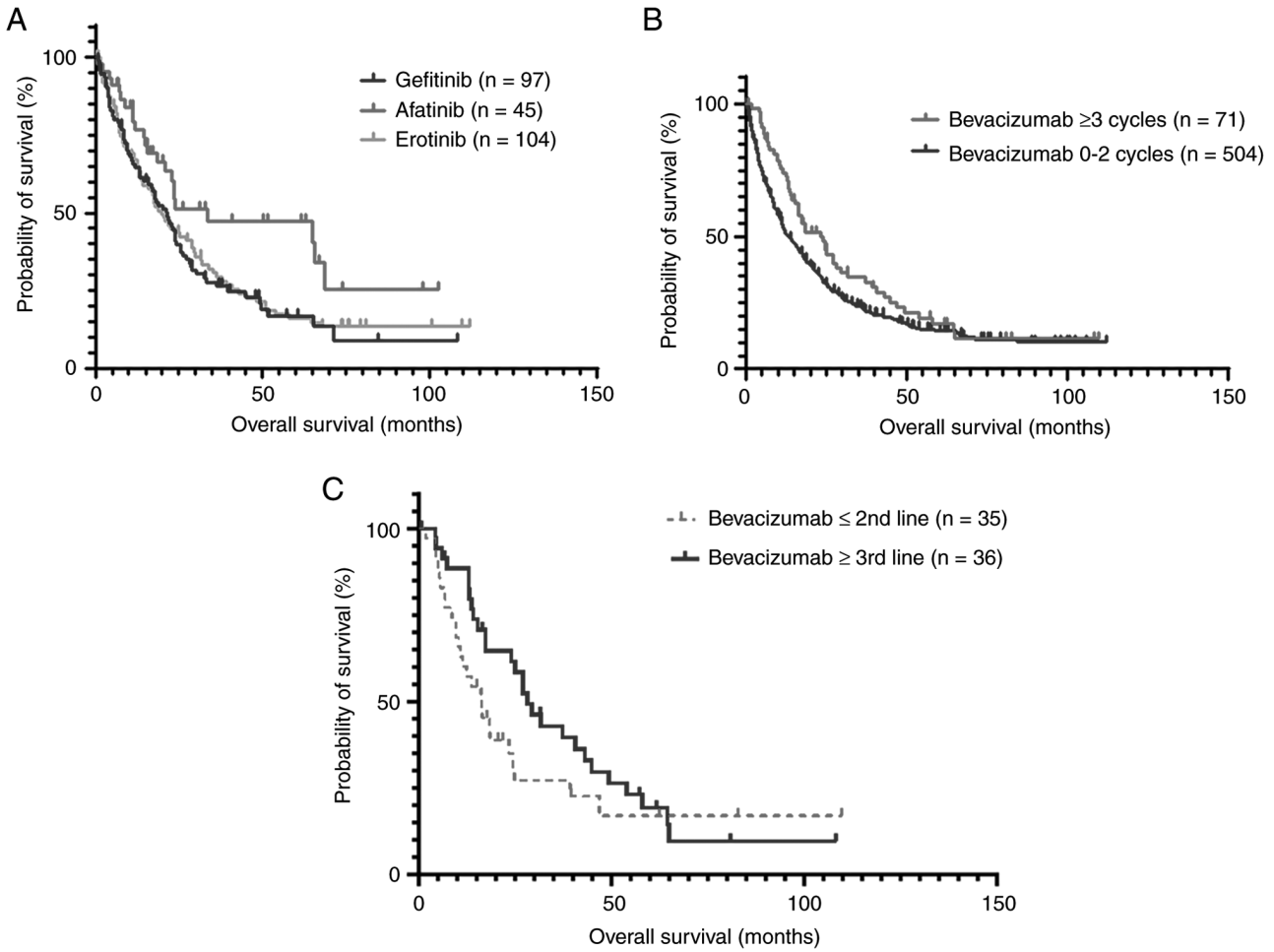

Figure 2. OS of patients with advanced lung adenocarcinoma according to therapeutic strategies. Kaplan-Meier plots for OS by (A) first-line targeted therapy in patients with epidermal growth factor receptor mutations, (B) bevacizumab treatment status and (C) bevacizumab treatment line. OS, overall survival.

Table IV. C-index of different machine learning models in survival analysis.

Machine learning models	Mean C-index	Standard deviation
EFGR variant patterns	0.614	0.007
Cox proportional hazards	0.695	0.021
Support vector machine	0.702	0.023
Random survival forest	0.689	0.020
Gradient boosting survival	0.683	0.022

All models were built on 70% training set and validated on 30% test set, with the C-index being the average result on the test set. C-index, concordance index; EGFR, epidermal growth factor receptor.

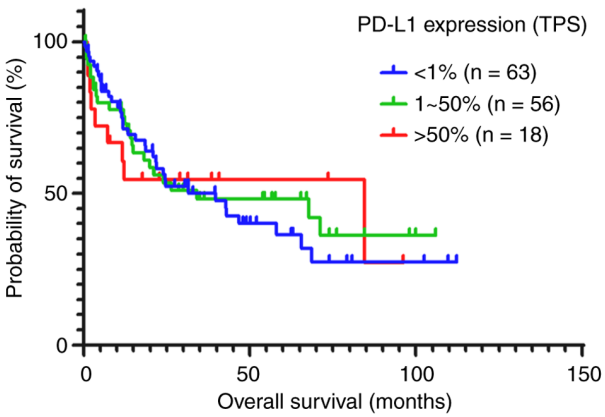

Figure 3. Association between PD-L1 expression (tumor proportion score) and overall survival in patients with advanced lung adenocarcinoma. PD-L1, programmed death-ligand 1.

period (median, 8.6 months), consistent with previous reports (15). Notably, most cases within this category involved compound mutations, 13 in total, including L858/T790M, L858R/L861Q and exon19 del/T790M. These EGFR compound mutations were revealed to be associated with a relatively long OS time, consistent with the findings of a previous study (16). Patients with the third most common rare mutation, the simple exon 21-point mutation L861Q, had a median OS of 50.8 months, and exhibited a favorable response to the second-generation TKI afatinib.

With respect to ALK-positive cases, only 10 of the 255 patients tested were found to harbor ALK rearrangements, accounting for 3.92% of the study cohort. Due to the small

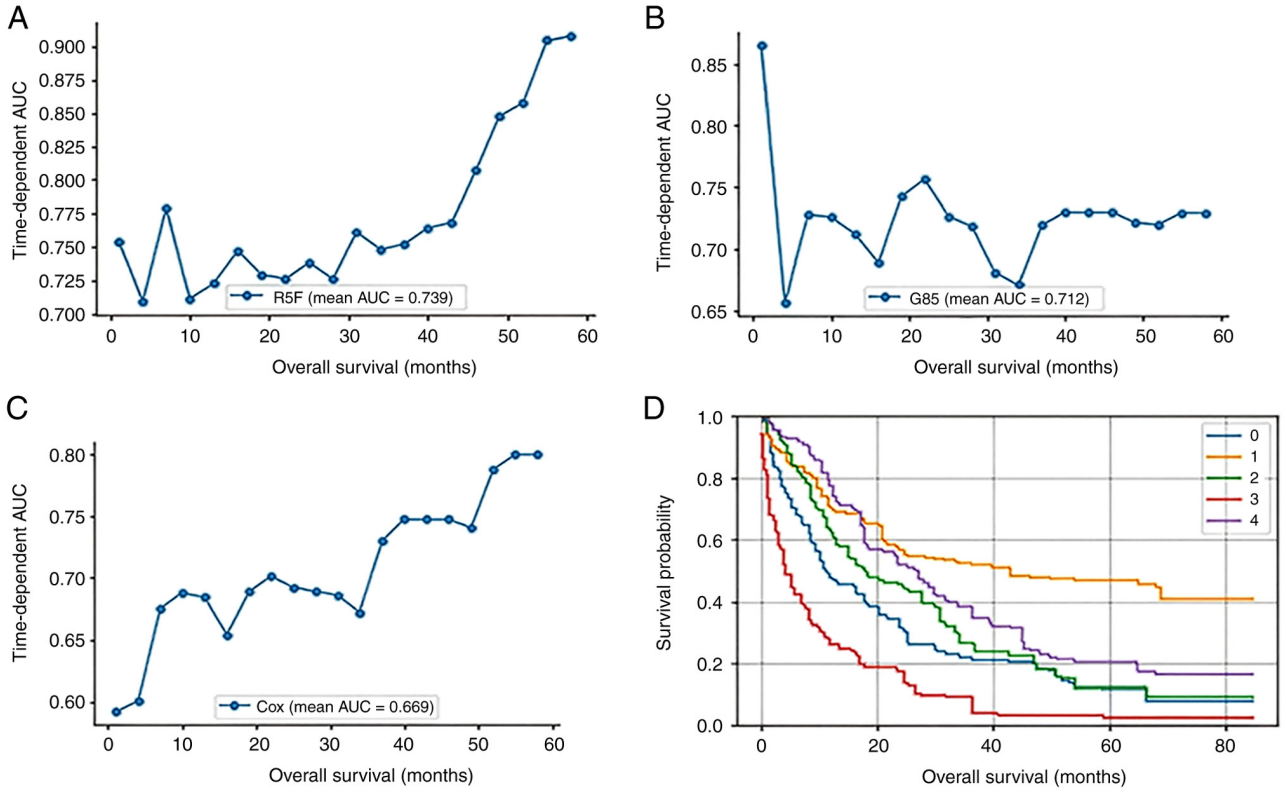

Figure 4. Machine learning-based survival analyses. Time-dependent C-index of (A) RSF, (B) GBS and (C) Cox proportional hazards models in a single representative subject. (D) Predicted survival probabilities for five representative patients using the RSF model; 0-4 are patient identification numbers. RSF, random survival forest; GBS, gradient boosting survival.

number of cases, definitive conclusions cannot be drawn; however, treated ALK-positive patients exhibited a long OS, which is consistent with general clinical observations and previously published studies (17,18).

Comparative analysis of first-line TKIs revealed a clear disparity in OS, with the OS of afatinib being longer than that of erlotinib and gefitinib. This finding provides insight into the relative efficacy of first-line TKIs, and may be used to inform treatment decisions and guide clinicians toward the most effective therapeutic options for patients with advanced adenocarcinoma.

Angiogenesis inhibitors such as bevacizumab have been widely used to treat ovarian cancer, and their incorporation into combination regimens is an important strategy for optimizing therapeutic outcomes. In EGFR-mutated metastatic NSCLC, the combination of bevacizumab and erlotinib has been shown to significantly improve PFS and overall response rates, but is also associated with high toxicity (6). Despite this efficacy, no improvement in OS has been reported for this combination (6,19,20). Bevacizumab is also used as an adjuvant treatment for lung adenocarcinoma, often in combination with other targeted therapies or chemotherapy. However, evidence regarding whether the inclusion of bevacizumab can prolong OS when used as a first-line therapy is inconsistent (21,22).

The present study showed that, regardless of EGFR mutation status, patients with lung adenocarcinoma who received more than three doses of bevacizumab had a significantly prolonged OS time. Further analysis suggested that the use of bevacizumab in the third or later treatment line did not significantly prolong OS compared with earlier use after accounting for potential hazard crossing. Therefore, the findings do not provide definitive evidence that adding anti-angiogenic therapy in later lines confers a greater survival benefit than is achieved when it is used in earlier lines. Notably, these findings may be influenced by selection bias, as patients who are able to receive third-line therapy may inherently have improved prognoses and longer OS. As a result, the analysis may not allow detection of a true treatment-line-dependent effect of bevacizumab.

Immunotherapy is an additional therapeutic option for the treatment of lung adenocarcinoma. Since immunotherapy was first used for melanoma, its use has gradually expanded to the treatment of various types of cancer. The expression of PD-L1 in tumor tissue is considered to inhibit the cytotoxic activity of immune cells against cancer cells and is commonly used to predict the response to immunotherapy (23). Increased PD-L1 expression on tumor-infiltrating cells has been associated with more aggressive tumor behavior (24). In studies of NSCLC, patients with high PD-L1 expression in the absence of EGFR mutation exhibited poorer OS than those with lower PD-L1 expression levels (25,26). In the present analysis, 137 patients underwent PD-L1 testing using the 22C3 monoclonal antibody, of whom 63 (46%) had PD-L1 <1%, 56 (41%) had PD-L1 between 1 and 50%, and only 18 (13%) had PD-L1 >50%. Among these groups, the PD-L1 >50% group appeared to have the longest median OS, although no statistically significant difference among the groups was observed. Variability in treatment approaches among the groups may have influenced

the observed outcomes. A previous study suggested that maintenance permetrexed chemotherapy was associated with improved progression-free survival and OS in patients with lung adenocarcinoma exhibiting high PD-L1 expression levels compared with those with low PD-L1 expression (27). Among the 18 patients with PD-L1 >50% in the present study, five harbored EGFR mutations (three had exon 19 deletions and two had the L858R mutation), while of the remaining 13 patients without EGFR mutations, only three received immunotherapy and the others were treated with permetrexed chemotherapy. Therefore, we hypothesize that the use of permetrexed chemotherapy to treat patients with PD-L1 >50% may explain the prolongation of OS.

The survival analysis of patients with lung cancer is complicated by the presence of numerous patient characteristics and clinical variables. To address this, machine learning models were used to process these high-dimensional variables and predict the risk of patients. A previous study has used machine learning models to investigate patients with early-stage (stage I to III) NSCLC, importing 127 features (28). By contrast, the present study focused on patients with advanced lung adenocarcinoma, included 28 features, and used the C-index to estimate the consistency between the predicted results and observed outcomes. A random forest-based prediction model was used to calculate the C-index at different time points. The C-index gradually increased after the observation period exceeded 3 years,. This may indicate that the patients underwent marked changes in clinical condition during the first 3 years of treatment, resulting in greater discrepancies between predicted and actual outcomes and consequently lower C-index values during this period. These findings suggest that clinicians should closely monitor changes in patient condition during the first 3 years of treatment. Different prediction models were selected according to the C-index at different time points to improve the prediction accuracy for patient survival.

The present study had several limitations. In Taiwan, osimertinib is not broadly used and was reimbursed only for second-line treatment in patients with the T790M mutation following failure of treatment with gefitinib, erlotinib, afatinib or dacomitinib from 2020 onwards. Consequently, OS outcomes associated with first-line osimertinib were not available in the database used in the current study. Although osimertinib has been shown to extend PFS as a first-line treatment for Taiwanese patients with advanced EGFR-mutated NSCLC, it has not exhibited a statistically significant OS advantage over second-generation EGFR-TKIs (29). In addition, immunotherapy and next-generation sequencing-based genetic testing have not been widely used or integrated into routine clinical practice for lung cancer.

In the era of machine learning, the present analysis was extended to predictive modeling by evaluating concordance statistics for OS based on diverse features within the patient cohort. This approach not only refines our understanding of prognostic factors but also contributes to the development of predictive tools that may support more accurate clinical decision-making. In summary, the present study aimed to elucidate the complex interrelationships among therapeutic modalities, prognostic markers and patient outcomes in advanced lung adenocarcinoma. By synthesizing evidence from a large-scale retrospective analysis, the findings of the study advance current knowledge of optimal treatment strategies and may pave the way for personalized interventions in this challenging clinical landscape.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The data generated in the present study may be requested from the corresponding author.

Authors' contributions

WTC designed the methodology, analyzed the data and wrote the original draft. SWC and SY participated in data analysis and revised the manuscript. IC was involved in conception and design of the analysis plan, critically revised the manuscript for important intellectual content, and provided substantive scientific input during manuscript amendments. SY participated in the coding and design of the methodology. WTC and SWC confirm the authenticity of all the raw data. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

This retrospective study was approved by the Institutional Review Board of Cheng-Hsin General Hospital [approval no. CHGH-IRB (922)111-01]. The requirement for informed consent was waived.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Jeon DS, Kim HC, Kim SH, Kim TJ, Kim HK, Moon MH, Beck KS, Suh YG, Song C, Ahn JS, *et al*: Five-year overall survival and prognostic factors in patients with lung cancer: Results from the Korean association of lung cancer registry (KALC-R) 2015. Cancer Res Treat 55: 103-111, 2023.
2. Osmani L, Askin F, Gabrielson E and Li QK: Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. Semin Cancer Biol 52: 103-109, 2018.
3. Greenhalgh J, Boland A, Bates V, Vecchio F, Dundar Y, Chaplin M and Green JA: First-line treatment of advanced epidermal growth factor receptor (EGFR) mutation positive non-squamous non-small cell lung cancer. Cochrane Database Syst Rev 3: CD010383, 2021.
4. Duma N, Santana-Davila R and Molina JR: Non-small cell lung cancer: Epidemiology, screening, diagnosis, and treatment. Mayo Clin Proc 94: 1623-1640, 2019.

5. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, Nicholson AG, Groome P, Mitchell A, Bolejack V, *et al*: The IASLC lung cancer staging project: Proposals for revision of the TNM stage groupings in the forthcoming (Eighth) Edition of the TNM classification for lung cancer. J Thorac Oncol 11: 39-51, 2016.
6. Motta-Guerrero R, Leon Garrido-Lecca A, Failoc-Rojas VE, Calle-Villavicencio A, Villacorta-Carranza R, Huerta-Collado Y, Torres-Mera A, Valladares-Garrido MJ, Rivera-Francia V, Carracedo C and Raez L: Effectiveness and safety of the bevacizumab and erlotinib combination versus erlotinib alone in EGFR mutant metastatic non-small-cell lung cancer: Systematic review and meta-analysis. Front Oncol 13: 1335373, 2024.
7. Zeng L, Cook RJ, Wen L and Boruvka A: Bias in progression-free survival analysis due to intermittent assessment of progression. Stat Med 34: 3181-3193, 2015.
8. Imai H, Kaira K and Minato K: Clinical significance of post-progression survival in lung cancer. Thorac Cancer 8: 379-386, 2017.
9. Li J, Zhu H, Sun L, Xu W and Wang X: Prognostic value of site-specific metastases in lung cancer: A population based study. J Cancer 10: 3079-3086, 2019.
10. Li Y, Wong M, Zhan L, Corke L, Brown MC, Cheng S, Khan K, Balatnaram K, Chowdhury M, Sabouhanian A, *et al*: Single organ metastatic sites in non-small cell lung cancer: Patient characteristics, treatment patterns and outcomes from a large retrospective Canadian cohort. Lung Cancer 192: 107823, 2024.
11. Tseng CH, Chiang CJ, Tseng JS, Yang TY, Hsu KH, Chen KC, Wang CL, Chen CY, Yen SH, Tsai CM, *et al*: EGFR mutation, smoking, and gender in advanced lung adenocarcinoma. Oncotarget 8: 98384-98393, 2017.
12. Hsu CH, Tseng CH, Chiang CJ, Hsu KH, Tseng JS, Chen KC, Wang CL, Chen CY, Yen SH, Chiu CH, *et al*: Characteristics of young lung cancer: Analysis of Taiwan's nationwide lung cancer registry focusing on epidermal growth factor receptor mutation and smoking status. Oncotarget 7: 46628-46635, 2016.
13. Kim TH, Choi JH, Ahn MS, Lee HW, Kang SY, Choi YW, Koh YW and Sheen SS: Differential efficacy of tyrosine kinase inhibitors according to the types of EGFR mutations and agents in non-small cell lung cancer: A real-world study. BMC Cancer 24: 70, 2024.
14. Hou J, Li H, Ma S, He Z, Yang S, Hao L, Zhou H, Zhang Z, Han J, Wang L and Wang Q: EGFR exon 20 insertion mutations in advanced non-small-cell lung cancer: Current status and perspectives. Biomark Res 10: 21, 2022.
15. Beau-Faller M, Prim N, Ruppert AM, Nanni-Metellus I, Lacave R, Lacroix L, Escande F, Lizard S, Pretet JL, Rouquette I, *et al*: Rare EGFR exon 18 and exon 20 mutations in non-small-cell lung cancer on 10 117 patients: A multicentre observational study by the French ERMETIC-IFCT network. Ann Oncol 25: 126-131, 2014.
16. Rossi S, Damiano P, Toschi L, Finocchiaro G, Giordano L, Marinello A, Bria E, D'Argento E and Santoro A: Uncommon single and compound EGFR mutations: Clinical outcomes of a heterogeneous subgroup of NSCLC. Curr Probl Cancer 46: 100787, 2022.
17. Lin CW, Huang KY, Lin CH, Hou MH and Lin SH: Diverse clinical outcomes for the EGFR-mutated and ALK-rearranged advanced non-squamous non-small cell lung cancer. Oncol Lett 29: 125, 2025.
18. Shimamura SS, Shukuya T, Asao T, Hayakawa D, Kurokawa K, Xu S, Miura K, Mitsuishi Y, Tajima K and Shibayama R, *et al*: Survival past five years with advanced, EGFR-mutated or ALK-rearranged non-small cell lung cancer-is there a 'tail plateau' in the survival curve of these patients? BMC Cancer 22: 323, 2022.
19. Colomban O, Tod M, Peron J, Perren TJ, Leary A, Cook AD, Sajous C, Freyer G and You B: Bevacizumab for newly diagnosed ovarian cancers: Best candidates among high-risk disease patients (ICON-7). JNCI Cancer Spectr 4: pkaa026, 2020.
20. Tewari KS, Burger RA, Enserro D, Norquist BM, Swisher EM, Brady MF, Bookman MA, Fleming GF, Huang H, Homesley HD, *et al*: Final overall survival of a Randomized trial of bevacizumab for primary treatment of ovarian cancer. J Clin Oncol 37: 2317-2328, 2019.
21. Reck M, Von Pawel J, Zatloukal P, Ramlau R, Gorbounova V, Hirsh V, Leighl N, Mezger J, Archer V, Moore N, *et al*: Overall survival with cisplatin-gemcitabine and bevacizumab or placebo as first-line therapy for nonsquamous non-small-cell lung cancer: Results from a randomised phase III trial (AVAiL). Ann Oncol 21: 1804-1809, 2010.
22. Sato H, Nagashima H, Akiyama M, Ito T, Hashimoto T, Saikawa H, Utsumi Y and Maemondo M: Analysis of bevacizumab treatments and metastatic sites of lung cancer. Cancer Treat Res Commun 26: 100290, 2021.
23. Chen G, Huang AC, Zhang W, Zhang G, Wu M, Xu W, Yu Z, Yang J, Wang B, Sun H, *et al*: Exosomal PD-L1 contributes to immunosuppression and is associated with anti-PD-1 response. Nature 560: 382-386, 2018.
24. Ren M, Dai B, Kong YY, Lv JJ and Cai X: PD-L1 expression in tumour-infiltrating lymphocytes is a poor prognostic factor for primary acral melanoma patients. Histopathology 73: 386-396, 2018.
25. Zhou ZJ, Zhan P and Song Y: PD-L1 over-expression and survival in patients with non-small cell lung cancer: A meta-analysis. Transl Lung Cancer Res 4: 203-208, 2015.
26. Tang Y, Fang W, Zhang Y, Hong S, Kang S, Yan Y, Chen N, Zhan J, He X, Qin T, *et al*: The association between PD-L1 and EGFR status and the prognostic value of PD-L1 in advanced non-small cell lung cancer patients treated with EGFR-TKIs. Oncotarget 6: 14209-14219, 2015.
27. Qin Y, Jiang L, Yu M, Li Y, Zhou X, Wang Y, Gong Y, Peng F, Zhu J, Liu Y, *et al*: PD-L1 expression is a promising predictor of survival in patients with advanced lung adenocarcinoma undergoing pemetrexed maintenance therapy. Sci Rep 10: 16150, 2020.
28. Kinoshita F, Takenaka T, Yamashita T, Matsumoto K, Oku Y, Ono Y, Wakasu S, Haratake N, Tagawa T, Nakashima N and Mori M: Development of artificial intelligence prognostic model for surgically resected non-small cell lung cancer. Sci Rep 13: 15683, 2023.
29. Chen HY, Chen CH, Liao WC, Lin YC, Chen HJ, Hsia TC, Cheng WC and Tu CY: Optimal first-line treatment for EGFR-mutated NSCLC: A comparative analysis of osimertinib and second-generation EGFR-TKIs. BMC Pulm Med 24: 517, 2024.

Copyright © 2026 You et al. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.