Clinical and Genomic Biomarkers for Progression-Free Survival in Non-Small Cell Lung Cancer: A Machine Learning Approach

Owen Zhuang Sun*
California Academy of Mathematics and Science
Carson, California

Abstract

Lung cancer is the leading cause of cancer mortality in the United States, with non-small cell lung cancer (NSCLC) accounting for approximately 85% of cases. This study aims to identify clinical and genomic risk factors associated with progression-free survival (PFS) in advanced NSCLC patients. A cohort of 218 U.S. patients from the MSK MIND dataset was analyzed using three survival analysis models implemented in Python. The analysis revealed that *EGFR* and *STK11* driver mutations and elevated derived neutrophil-to-lymphocyte ratio (dNLR) levels were associated with increased hazard. Albumin levels were associated with a significant decrease in hazard. PD-L1 expression and tumor mutational burden (TMB) showed relatively modest protective effects. The Gradient Boosted Machine (GBM), a machine learning model for survival analysis, demonstrated the highest predictive capability with a C-index of 0.701, having better-than-random performance in the testing dataset. These findings highlight the critical role of specific clinical and genomic biomarkers in affecting NSCLC survival and improving the accuracy of survival predictions.

Keywords: non-small cell lung cancer, progression-free survival, survival analysis with censoring, Cox proportional hazards model, random survival forests, gradient boosting survival analysis

1 Introduction

1.1 Background

Lung cancer, the leading cause of cancer deaths, will account for an estimated 125,070 deaths in 2024 (Siegel, Giaquinto, and Jemal 2024). Non-small cell lung cancer (NSCLC) is the most common subtype of lung cancer, accounting for around 85% of lung cancer diagnoses (Chevallier et al. 2021). The primary risk factor for lung cancer development is smoking, either directly or through secondhand smoke, followed by a family history of lung cancer, and exposure to carcinogenic substances such as asbestos (Alexander, Kim, and Cheng 2020).

Progression-free survival (PFS), the time until NSCLC stage progression or death, is often used as an endpoint to assess therapy efficacy and identify risk factors associated with disease advancement or death (Alexander, Kim, and Cheng 2020). This study investigates several clinical and genomic risk factors associated with NSCLC progression.

1.2 Literature Review

In a retrospective claims-dataset analysis of 1,741 German advanced NSCLC patients observed between 2011 and 2016, Hardtstock et al. (2020) found that treatment with targeted therapy and/or immunotherapy was associated with a decrease in hazard — as performed by a Cox proportional

*osun342@gmail.com

hazards model — and significantly higher overall survival than chemotherapy alone. Braghetto et al. (2022) used deep learning to first extract radiomics features for the prediction of 2-year overall NSCLC survival and then found moderate discriminatory capability across several models, including Cox proportional hazards and random survival forest models. Germer et al. (2024) found that Cox proportional hazards and random survival forest models perform comparably and have moderate discriminatory capability given data on sex, age, histology, and cancer stage from a German cancer registry. Using electronic health records data, Li et al. (2022) found that Cox proportional hazards gradient-boosted decision trees had moderate discriminatory capability and outperformed other models in predicting PFS and overall survival, including Cox proportional hazards model, accelerated failure time model, and survival support vector machines, and DeepSurv (a deep-learning survival framework). Furthermore, important features from the GBM included programmed cell death-ligand 1 expression (PD-L1), Eastern Cooperative Oncology Group performance score (ECOG), and serum albumin. Cramer-van der Welle et al. (2021) compared real-world survival outcomes with clinical trial results for immunotherapy using Kaplan-Meier and Cox proportional hazards model. Cramer-van der Welle and colleagues (2021) found that although immunotherapy PFS results are similar between clinical trial results and real-world outcomes, pembrolizumab overall survival was significantly shorter in the real world than in the trial and was associated with increased hazard compared to the clinical trial data.

In general, the Cox proportional hazards model is more widely used to quantify hazard ratios between treatment and control groups in clinical trials. For instance, Soria et al. (2018) demonstrated that in untreated *EGFR*-mutant NSCLC patients, osimertinib, a tyrosine kinase inhibitor, was associated with a statistically significant reduction in hazard compared to standard *EGFR* tyrosine kinase inhibitors.

1.3 Data Description

The data in this study originates from the Memorial Sloan Kettering Cancer Center's MIND (Multimodal Integration of Data) initiative (Vanguri et al. 2022). The publicly available data were accessed via cBioPortal (Cerami et al. 2012).

The data include 247 of the Center's patients with advanced NSCLC who received PD-(L)1-blockade therapy between 2014 and 2019 (Vanguri et al. 2022). Following the removal of patients lacking data on the variables of interest, time-to-event analysis for progression-free survival was performed on 218 patients. Categorical data were one-hot encoded for analysis. Degenerate parameters (*ALK*, *RET*, and *ROS1* drivers) were removed. Drivers (*MET*, *BRAF*, and *ARID1A* drivers) without any statistically significant impacts on hazard were dropped. Immunotherapy treatments (e.g., pembrolizumab, atezolizumab, and nivolumab) were excluded from analysis due to being highly collinear with clinically-reported PD-L1 score; PD-L1 score is used as a biomarker for determining if and which immunotherapy are provided to patients (Vanguri et al. 2022).

Clinical biomarkers investigated in this article include ECOG (Eastern Cooperative Oncology Group) performance status, derived neutrophil-to-lymphocyte ratio (dNLR), pack-year history, and albumin. ECOG performance status is a 5-point score to evaluate the self-care capability of patients and is used in clinical decision-making on systemic treatment (Azam et al. 2019). The dNLR quantifies the balance between neutrophils and lymphocytes. Neutrophils, key components of the innate immune system, mediate inflammation and are typically elevated during chronic inflammation (Yang et al. 2021). In a proinflammatory environment, immature neutrophils can be released from the bone marrow, rapidly increasing neutrophil count. Conversely, lymphocytes, part of the adaptive immune system, tend to decrease during chronic inflammation. As a result, dNLR serves as a potential biomarker of systemic inflammation and has been linked to poorer survival outcomes in NSCLC (Yang et al. 2021). Albumin is an abundant circulating protein produced in the liver used as a biomarker for nutritional status. It may also inform on systemic inflammation as albumin synthesis is reduced by inflammation-inducing factors, notably tumor necrosis factor (TNF) (Zhang et al. 2023). Smoking, quantified by pack-year history — the number of daily packs smoked multiplied by the number of years smoked, correlates with increased incidence of NSCLC and changes to the tumor microenvironment (Zhao et al. 2021). A meta-analysis by Zhao et al. (2021) also reported

higher response rates to immunotherapies among smokers than non-smokers, along with greater overall survival and progression-free survival.

Genomic biomarkers are commonly used to assess eligibility for immune checkpoint inhibitors and targeted therapies. Key biomarkers include PD-L1 expression and, historically, tumor mutational burden (TMB) (Ettinger et al. 2022). PD-L1 expression inhibits CD8+ T lymphocyte-induced apoptosis, protecting the tumor from the immune system (Yu et al. 2016). PD-L1 is characterized by PD-L1 score: the proportion of a tumor sample's expression of PD-L1 protein. A PD-L1 score above 50% was a key criterion in the KEYNOTE-024 trial, which demonstrated significantly improved progression-free survival in advanced NSCLC patients treated with pembrolizumab immunotherapy (Reck et al. 2016). TMB, measured using the MSK-IMPACT assay, quantifies the number of mutations present in the tumor genome and serves as an additional biomarker for predicting response to immunotherapy response (Vanguri et al. 2022).

Certain mutations in cancer driver genes can result in or aid in tumor growth and proliferation. Several of these mutations are also clinically actionable with targeted therapy (Ettinger et al. 2022). Two of these driver oncogenes are *EGFR* (HER1 or *ERBB1*) and *ERBB2* (HER2 or HER2/neu), both members of the human epidermal growth factor (HER/ERBB) family of receptor tyrosine kinases that trigger cell replication (Chevallier et al. 2021). Activating mutations in the tyrosine kinase domain of *EGFR* or *ERBB2* can lead to constitutive activation of their respective signaling pathways without ligand binding, causing uncontrolled cell proliferation and, ultimately, contributing to oncogenesis (Chevallier et al. 2021). Approximately 19% of NSCLC patients have an *EGFR* driver mutation while 1 to 3% has an *ERBB2* driver mutation (Chevallier et al. 2021). *EGFR* mutations are clinically actionable with targeted therapies, such as osimertinib (Ettinger et al. 2022). Previously, anti-HER2 therapies showed no clear benefit in NSCLC with overactive HER2 — the receptor tyrosine kinase encoded by *ERBB2* (Chevallier et al. 2021). However, in August 2022, the U.S. Food and Drug Administration (FDA) granted accelerated approval for fam-trastuzumab deruxtecan-nxki for treating metastatic NSCLC with an *ERBB2* driver mutation (U.S. Food and Drug Administration 2022).

STK11 is a driver tumor suppressor gene, and approximately 10% of NSCLC patients have an *STK11* mutation (Malhotra et al. 2022). *STK11* inactivation contributes to uncontrolled cell proliferation and metabolic changes in the tumor, leading to an altered tumor microenvironment hostile to cytotoxic CD8+ T lymphocytes (Malhotra et al. 2022) and reducing immune surveillance. This reduction is associated with lower tumor PD-L1 expression, leading *STK11* mutations to be the most significant molecular factor currently known to drive PD-L1 immunotherapy resistance in NSCLC (Malhotra et al. 2022). Consequently, *STK11*-mutant NSCLC is associated with worse survival (Malhotra et al. 2022).

1.4 Time to Event Analysis

Time-to-event analysis, also known as survival analysis, investigates the duration of time until an event of interest, such as death, occurs and the factors associated with event occurrence and timing. The event of interest in this study is cancer progression or cancer-related death, whichever occurs first, to investigate progression-free survival (PFS).

Particularly in cohort studies — such as the one analyzed in this report, patients may drop out of the study or be lost to follow-up. These patients are considered right-censored (referred to as simply "censored" in this study) as the exact time until the event is unknown; however, it *is* known that they did not experience the event up until the time they were last observed. Thus, censored data are still valuable despite not providing the exact time of the event.

Survival analysis, incorporating censored data, seeks to estimate and model several key functions: the survival function, $S(t)$; hazard function, $h(t)$; and cumulative hazard function, $H(t)$. These functions are defined as follows. Assume T is the random variable of the time to disease progression or death for a patient, and let $f(t)$, $t \geq 0$, denote the probability density function (pdf) of T and $F(t) = \mathbb{P}(T \leq t) = \int_0^t f(x)\,dx$, $t \geq 0$, be the cumulative distribution function (cdf) of T. The survival function, $S(t)$, is the probability that the patient survives without progression until time t:

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(x)dx = 1 - F(t), t \geq 0.$$

The hazard function $h(t)$ is defined by the formula $h(t) = \frac{f(t)}{S(t)}$, $t \geq 0$, and can be interpreted as the instantaneous rate of disease progression or death, given that the patient survived up to time t. This can be derived by writing

$$h(t) = \mathbb{P}(T < t + dt \,|\, T > t) = \frac{\mathbb{P}(t < T < t + dt)}{\mathbb{P}(T > t)} \approx \frac{f(t)dt}{S(t)} = h(t)dt.$$

The cumulative hazard function $H(t)$ is the cumulative risk of experiencing the event until time t:

$$H(t) = \int_0^t h(x)\, dx.$$

1.5 Organization

This paper is organized into four primary sections: Introduction, Methods, Results, and Discussion. In the Methods section, the three survival model frameworks evaluated are discussed: the Cox proportional hazards model, random survival forests, and gradient-boosted machines. The Results section covers the overall performance of each model and the important and/or statistically significant parameters determined by each model. The Discussion section connects the parameters isolated by the models to their clinical contexts and discusses limitations and next steps. The significance level is set to be 0.05.

2 Methods

2.1 Cox Proportional Hazards Model

A Cox proportional hazards model, also known as a Cox model, models the hazard function given $x_1, ..., x_m$ covariates and assumes a baseline risk of $h_0(t)$ at time t:

$$h(t, x_1, ..., x_m, \beta_1, ..., \beta_m) = h_0(t) \exp(\beta_1 \, x_1 \, + \, \cdots \, + , \beta_m \, x_m).$$

Notably, the model requires covariates to remain constant over time. Parameter estimation can be performed using the maximum-likelihood approach, allowing for the determination of the most probable values given the data.

2.2 Random Survival Forests

As introduced by Ishwaran et al. (2008), random survival forests (RSF) is an ensemble learning technique using multiple decision trees as base learners, where the predicted hazard is the average of the hazard predicted by terminal nodes. An advantage of RSF is the ability of the trees to learn non-linear relationships and interactions between covariates.

The algorithm is as follows:

1. Randomly select B bootstrap samples—selecting samples with replacement—from the dataset.
2. To grow a survival tree for each sample, use the log-rank test statistic to determine the parameter that maximizes survival difference. The node is split according to this parameter such that the survival difference is as large as possible between the two daughter nodes.

The log-rank test statistic is calculated as follows. With the split between the two groups (left and right daughter nodes) indicated by superscripts 1 and 2, the observed number of events at time t_i are O_{1i} and $O_{2i},\ i = 1, \ldots, k$. The expected number of events for group 1 is

$$\mathbb{E}_i^1 = \frac{N_{1i}}{N_i} O_i$$

where N_{1i} represents the number of patients at risk in group 1 at time t_i, and N_i is the total number of patients at risk at time t_i. The variance of O_i^1 is given by:

$$\mathbb{V}ar(O_i^1) = \frac{N_{1i}\, N_{2i}\, (N_i - O_i)\, O_i}{N_i^2\, (N_i\, -\, 1)}.$$

The log-rank test statistic is defined by the formula

$$L = \frac{\sum_{i=1}^{k} (O_{1i} - \mathbb{E}_i^1)}{\sqrt{\sum_{i=1}^{k}\ \mathbb{V}ar(O_i^1)}}$$

The parameter with the largest log-rank test statistic is used to split the node.

3. Continue this process to develop the tree until a threshold for the number of events in each terminal node is reached. Theoretically, the minimum threshold of events per terminal node is one event. Practically, this threshold is set above 1 to reduce overfitting.

4. Compute the cumulative hazard function (CHF) for each terminal node, then average these CHFs across terminal nodes to obtain the ensemble's CHF.

Each case within the same node has the same CHF as others in the same node. The CHF is based on the Nelson-Aalen cumulative hazard estimator. For each node, the CHF is estimated as follows:

$$\hat{H}(t)\ =\ \sum_{0\,\leq\, t_i\, \leq\, t} \frac{O_{1i}}{N_{1i}}.$$

The ensemble CHF is estimated by taking the average of the CHFs outputted by each terminal node. Given B terminal nodes, the final estimate of the CHF computed as:

$$\hat{H}(t \mid x)\ =\ \frac{1}{B} \sum_{b=1}^{B} \hat{H}_b(t \mid x).$$

Note: The estimated CHF can be converted to an estimated survival function using the following relation:

$$\hat{S}(t \mid x)\ =\ \exp\left(-\hat{H}(t \mid x)\right).$$

5. The concordance index (C-index) measures the predictive accuracy of a RSF. This index evaluates how well the model ranks survival times, assigning higher predicted risks to shorter survival times.

Below we outline how the C-index is computed. We look at all possible pairs of cases in the dataset, but we exclude any pairs where: (1) one case has a censored survival time shorter than the other case's survival time, or (2) both cases have the same survival time unless at least one of them experienced the event (i.e., wasn't censored). For each eligible pair (i, j), with survival times T_i and T_j:

- If $T_i \neq T_j$ (the survival times differ), add 1 if the model predicts a worse outcome for the case with the shorter survival time.

- If $T_i = T_j$ and both cases experienced the events, add 1 if the predicted outcomes are equal; otherwise, add 0.5 if the predictions differ.

- If $T_i = T_j$ but one case experienced the event and the other was censored, add 1 if the model predicts a worse outcome for the case with the event; otherwise, add 0.5 if the predictions differ.

The C-index is the total of these scores divided by the total number of eligible pairs. A C-index of 0.5 suggests that the model's predictions are no better than random chance, while a 1 indicates perfect prediction.
To estimate the parameters of the model, the algorithm repeatedly builds decision trees on random samples of the data and takes averages across these trees to stabilize estimates.

2.3 Gradient Boosting Machines

Gradient boosting machines (GBM) are another ensemble learning technique for survival analysis that uses regression trees as base learners to optimize based on the partial log-likelihood function for the Cox proportional hazards model. The partial log-likelihood function has the form:

$$\ell(\beta) \;=\; \sum_{i:\delta_i=1} \left(\mathrm{x}_i^\top \beta - \log \left(\sum_{j \in R(t_i)} e^{\mathrm{x}_j^\top \beta} \right) \right).$$

Here δ_i is the event indicator, that is, $\delta_i = 1$ if the event occurs, and $\delta_i = 0$ if the observation is censored, $\mathrm{x}_i^\top = \big(x_{i1},\, x_{i2},\, \cdots,\, x_{ip}\big)$ is a row-vector with covariates for case i, and $\beta = \big(\beta_1,\, \beta_2,\, \cdots,\, \beta_p\big)^\top$ is a column-vector of coefficients. Thus, $\mathrm{x}_i^\top \beta = x_{i1} \cdot \beta_1 + x_{i2} \cdot \beta_2 + \cdots + x_{ip} \cdot \beta_p$. Additionally, the hazard ratios $e^{x_j^\top \beta}$ are summed across all cases in the risk set at time t_i, denoted as $R(t_i)$ in the formula above.

To maximize the partial log-likelihood function, gradient descent in function space — also known as gradient boosting — is employed. While gradient descent traditionally optimizes parameters, gradient boosting applies it to functions, as GBMs create an ensemble of decision trees, effectively an ensemble of functions. In this context, the negative partial log-likelihood serves as the loss function. The model fit is improved by using the negative gradient of this loss to maximize the partial log-likelihood:

$$g_i \;=\; -\frac{\partial(-\ell(\beta))}{\partial\beta} \;=\; \frac{\partial(\ell(\beta))}{\partial\beta} \;=\; \delta_i \left(\mathrm{x}_i - \frac{\sum_{j \in R(T_i)} \mathrm{x}_j e^{\mathrm{x}_j^\top \beta}}{\sum_{j \in R(T_i)} e^{\mathrm{x}_j^\top \beta}} \right).$$

For each $(m+1)$st iteration, the model adds an additional fitted regression tree, $h_m(x)$, weighted by learning rate η:

$$F_{m+1}(x) \;=\; F_m(x) \;+\; \eta \cdot h_m(x).$$

The final model is expressed as follows:

$$F_M(x) \;=\; \sum_{m=1}^{M} \eta \cdot h_m(x).$$

Here, M denotes the number of boosting rounds taken to build the ensemble model.

3 Results

3.1 Cohort Clinical Characteristics

Demographic, clinical, and genomic characteristics of the advanced NSCLC cohort (n = 218) are summarized in Figure 1. The median age was 67.5 years (range: 38 to 88 years), and the median pack-years smoked was 28 (range: 0 to 165). Regarding driver mutations, there were 21 *EGFR*-mutant cases, 16 *ERBB2*-mutant cases, and 42 *STK11*-mutant cases, with mutations being non-exclusive. The median clinically reported PD-L1 score was 0 (range: 0 to 100), and approximately 26% of patients had PD-L1 scores over 50. The frequency of immunotherapy usage is provided in the Technical Resources (5).

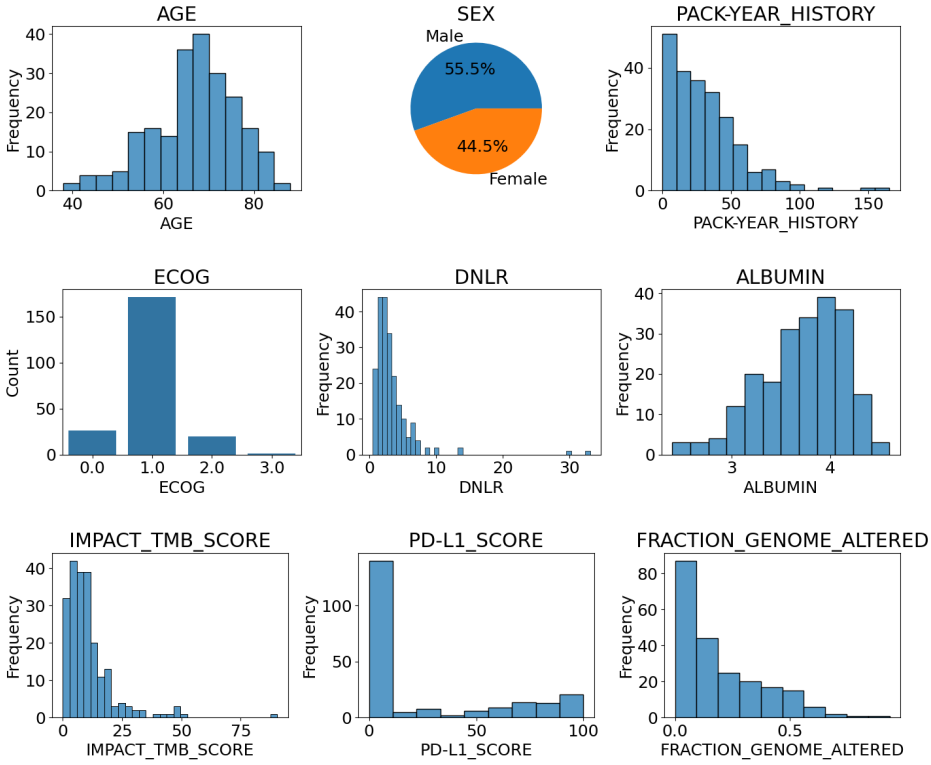

Figure 1: Cohort Demographic, Clinical, and Genomic Characteristics

The median progression-free survival was 2.7 months (range: 0.2 to 49.1 months). The Kaplan-Meier estimate of progression-free survival is displayed in Figure 2.

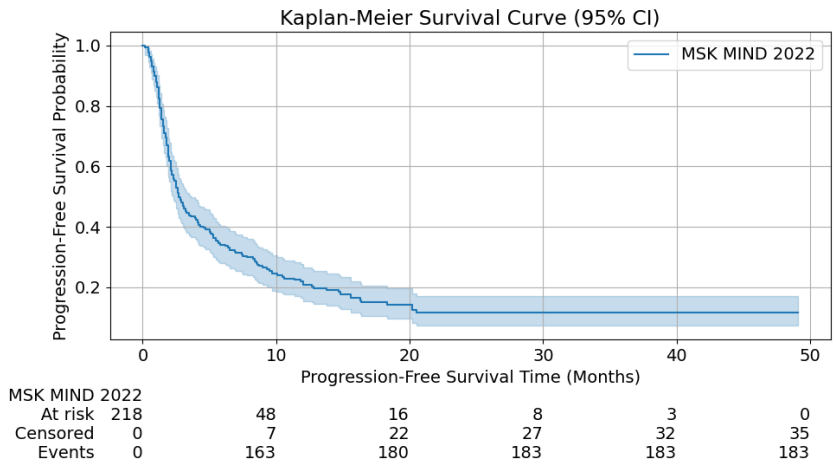

Figure 2: Kaplan-Meier Estimation of Overall Progression-Free Survival with 95% CI

3.2 Cox Proportional Hazards Model

Using the Cox proportional hazards model to fit main effects for ease of interpretation, clinical and genomic parameters were analyzed to evaluate the impacts of each parameter on PFS (Table 1). The concordance index (C-index) was 0.688 on the 20% test set.

The derived neutrophil-to-lymphocyte ratio (dNLR) was associated with a statistically significant increase in hazard (HR: 1.06, 95% CI: 1.0-1.13, p-value: 0.05), indicating a slight increase in risk of progression or death. The PD-L1 score was associated with a statistically significant but small decrease in hazard (HR: 0.99, 95% CI: 0.98-1.0, p-value: ≤ 0.001). Albumin showed the largest decrease in hazard (HR: 0.36, 95% CI: 0.23-0.56, p-value: ≤ 0.001).

On the genomic side, two driver mutations, *EGFR* and *STK11*, along with the TMB score, were linked to statistically significant impacts on hazard. Kaplan-Meier survival curves for *EGFR*- and *STK11*-mutant cases are provided in the Technical Resources (5). An *EGFR* mutation was associated with the largest increase in hazard (HR: 3.61, 95% CI: 2.01-6.5, p-value: ≤ 0.001). A mutation in the *STK11* tumor suppressor gene was associated with a statistically significant increase in hazard (HR: 1.73, 95% CI: 1.08-2.76, p-value: 0.022). Impact TMB score showed a significant reduction in hazard (HR: 0.96, 95% CI: 0.94-0.99, p-value: 0.002).

Table 1: Hazard Ratios on Clinical and Genomic Parameters (C-index: 0.688)

Covariate	Hazard Ratio	p
EGFR Driver	3.61 (2.01-6.5)	$\leq$ **0.001**
STK11 Driver	1.73 (1.08-2.76)	**0.022**
ERBB2 Driver	1.36 (0.72-2.58)	0.345
ECOG	1.29 (0.92-1.81)	0.138
MSI Score	1.07 (0.99-1.16)	0.073
dNLR	1.06 (1.0-1.13)	**0.05**
Age	1.0 (0.98-1.02)	0.926
Pack-Year History	1.0 (0.99-1.0)	0.334
Clinically Reported PD-L1 Score	0.99 (0.98-1.0)	$\leq$ **0.001**
Impact TMB Score	0.96 (0.94-0.99)	**0.002**
Is Female	0.91 (0.64-1.29)	0.584
Fraction Genome Altered	0.85 (0.35-2.07)	0.719
Albumin	0.36 (0.23-0.56)	$\leq$ **0.001**

3.3 Random Survival Forests

The hyperparameter-tuned RSF achieved a concordance index (C-index) of 0.691 on the 20% test set, slightly higher than the C-index of the main effects Cox model (C-index: 0.688).

Using permutation importance, features were ranked by their importance (Figure 3). The top six features were further analyzed with a Cox proportional hazards model to determine their hazard ratios (Table 2). The RSF-feature-selected Cox model achieved a C-index of 0.693 which slightly outperformed the main effects Cox model of clinical and genomic features (C-index: 0.688), suggesting that the RSF successfully performed feature selection.

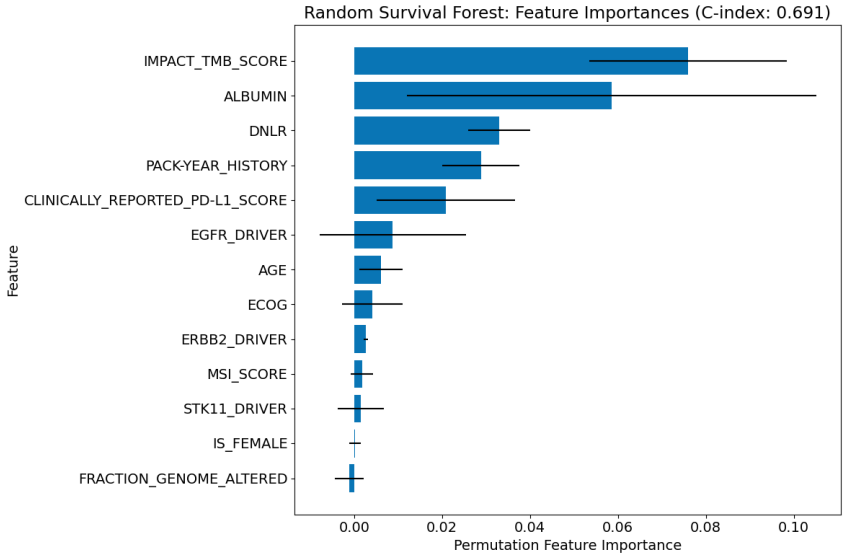

Figure 3: Random Survival Forest: Features by Permutation-Based Feature Importance

Similar to the main effects Cox proportional hazards model, several parameters selected by RSF were also found to be statistically significant (Table 2). Notably, the *EGFR* driver mutation was associated with an increased hazard (HR: 3.05, p-value: $\leq$ 0.001). An increased dNLR was also linked to a higher hazard, consistent with the main effects Cox model (HR: 1.09, p-value: 0.007).

PD-L1 and TMB scores were associated with a small decrease in hazard (Clinically Reported PD-L1 Score: HR: 0.99, p-value: ≤ 0.001; Impact TMB score: HR: 0.97, p-value: 0.003). Albumin was associated with the largest reduction in hazard (HR: 0.38, p-value: ≤ 0.001), aligning with the findings from the main effects Cox model. The statistically significant RSF-selected features also showed significance in the main effects Cox model, with similar hazard ratios.

Table 2: Hazard Ratios on RSF-Selected Clinical Parameters (C-index: 0.693)

Covariate	Hazard Ratio	p
EGFR Driver	3.05 (1.75-5.32)	$\leq$ **0.001**
dNLR	1.09 (1.02-1.15)	**0.007**
Pack-Year History	1.0 (0.99-1.0)	0.397
Clinically Reported PD-L1 Score	0.99 (0.98-0.99)	$\leq$ **0.001**
Impact TMB Score	0.97 (0.95-0.99)	**0.003**
Albumin	0.38 (0.26-0.56)	$\leq$ **0.001**

3.4 Gradient Boosting Machines

The gradient boosting machine (GBM), with regression trees as base learners and the Cox proportional hazards model's partial likelihood for loss, had a C-index of 0.701 on the 20% test set, higher than the main effects Cox and RSF models. Features were ranked by their importance using impurity-based feature importanceFigure 4.

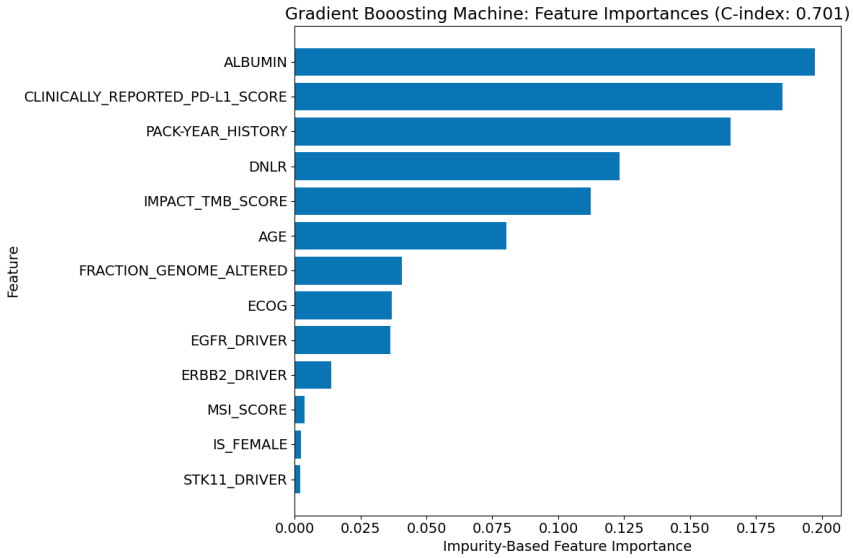

Figure 4: Gradient Boosting Machine: Features by Impurity-Based Feature Importance

The six leading features identified by GBM were further analyzed using a Cox proportional hazards model to assess their hazard ratios, resulting in a C-index of 0.669 (Table 3). The selected features were largely similar to those identified by the RSF, with the notable inclusion of age and the exclusion of the *EGFR* driver mutation.

The derived neutrophil-to-lymphocyte ratio (dNLR) was the only parameter associated with a statistically significant increase in hazard (HR: 1.08, p-value: 0.019). Consistent with previous models, PD-L1 and TMB scores exhibited statistically significant but small decreases in hazard

(Clinically Reported PD-L1 Score: HR: 0.99, p-value: 0.001; Impact TMB Score: HR: 0.97, p-value: 0.006). As observed in the main effects Cox models with all parameters and those selected by RSF, albumin was associated with a statistically significant decrease in hazard (HR: 0.39, p-value: ≤ 0.001).

Table 3: Hazard Ratios on GBM-Selected Clinical Parameters (C-index: 0.669)

Covariate	Hazard Ratio	p
dNLR	1.08 (1.01-1.15)	**0.016**
Age	1.0 (0.98-1.02)	0.873
Pack-Year History	0.99 (0.99-1.0)	0.216
Clinically Reported PD-L1 Score	0.99 (0.98-0.99)	$\leq$ **0.001**
Impact TMB Score	0.97 (0.95-0.99)	**0.006**
Albumin	0.39 (0.26-0.58)	$\leq$ **0.001**

4 Discussion

Several key insights were identified by the survival models, including the impact of the *EGFR* and *STK11* driver mutations, dNLR, PD-L1, and TMB.

The *EGFR* driver mutation was associated with the largest increase in hazard. At the start of the study period, the first line treatments for *EGFR*-mutant NSCLC were gefitinib and erlotinib, followed by afatinib and dacomitinib in the second-line (Chevallier et al. 2021). However, these therapies were shown to lead to systemic resistance, including resistance driven by T790M secondary mutations or activation of other *EGFR* pathways (Chevallier et al. 2021). As the cohort studied in this report composed of advanced NSCLC patients in whom *EGFR* was found to be linked with the largest increase in hazard, all patients with a *EGFR* driver mutation went through multiple lines of therapy and potentially developed systemic resistance to *EGFR* targeted therapy, ultimately leading to progression or death. However, in 2017, a new drug, osimertinib, was found to be efficacious in patients who had progressed through first- and second-line treatments and who had developed T790M secondary mutation in the AURA3 trial (Mok et al. 2017). As a result, osimertinib was approved as a third-line treatment by the FDA. In April 2018, osimertinib became a first-line treatment for *EGFR*-mutant NSCLC, following the FLAURA trial, after receiving FDA approval as a first-line therapy (Soria et al. 2018). By this time, only two patients in the cohort with *EGFR*-mutant NSCLC may have benefited from osimertinib as a first-line treatment. As a result of the advancements in targeted therapy, the hazard observed in *EGFR*-mutant NSCLC may not be reflected in current clinical practice.

Notably, both dNLR and albumin, biomarkers for systemic inflammation (Zhang et al. 2023) (Yang et al. 2021), were significant. Systemic inflammation has broad implications on tumors by stimulating tumorigenesis, increasing mutagenesis, promoting angiogenesis, inhibiting the adaptive immune response, and thereby influencing tumor responses to therapy and leading to tumor growth (Yang et al. 2021).

Higher dNLR is associated with tumor-induced chronic inflammation, as elevated neutrophil levels secrete proinflammatory cytokines, while lymphocyte count is diminished. Further, the inflammatory microenvironment created by tumor cells stimulates neutrophils and perpetuates the tumor inflammatory microenvironment (Yang et al. 2021). In a meta-analysis by Yang et al. (2021), increased dNLR was found to predict poor PFS in European or American patients but not in Asian patients. The elevated hazard observed in this study likely also reflects the characteristics of the cohort studied.

Albumin was associated with the greatest reduction in hazard. Albumin levels may be reduced due to poor nutrition, which is associated with poor prognosis in patients with NSCLC, or due to

inflammation. Albumin synthesis is inhibited by TNF, a proinflammatory cytokine, and in systemic inflammation, TNF may contribute to lower albumin levels (Zhang et al. 2023). Albumin is also part of the Advanced Lung Cancer Inflammation Index and Prognostic Nutritional Index, both of which have been associated with PFS prediction (Zhang et al. 2023).

The two biomarkers used for immunotherapy eligibility, PD-L1, and TMB, were both found to be statistically significant. Although their hazard ratios may not seem highly protective at 0.99 (95% CI: 0.98-1.0) and 0.96 (95% CI: 0.94-0.99), respectively, it is important to note that PD-L1 ranged from 0% to 100% and TMB from 0 to 90.4 mutations per megabase. For instance, a patient with a very high PD-L1 expression level of 100% would have an estimated hazard that is approximately 36% (95% CI: 13.262% - 100%) that of baseline. These results suggest that patients with sufficient PD-L1 expression and/or TMB who are treated with immunotherapy may experience increased PFS, consistent with the clinical evidence for immunotherapy (Alexander, Kim, and Cheng 2020). However, it is important to note that the confidence interval for PD-L1 approaches 1, suggesting that the observed association, while significant, may not be strongly protective in all cases. This warrants careful consideration in clinical decision-making.

The three models demonstrated moderate predictive accuracy overall. The predictive capability could be enhanced by the incorporation of multiomic data, including pathology and radiology data. Cox proportional hazards model performance may be improved by including interaction terms rather than limiting the model to main effects. Additionally, a greater sample size could better empower the machine learning models used and potentially allow for both increased prediction accuracy and deeper insights with the application of deep learning approaches.

5 Technical Resources and Repository

Cox proportional hazards models and Kaplan-Meier survival curves were implemented using the *lifelines* library (Davidson-Pilon 2019). Random survival forests and gradient-boosted machines were implemented using the *scikit-survival* library (Pölsterl 2020). The code and figures used in this study are available at: https://github.com/osun24/mskmind-survival

6 Acknowledgement

I am deeply grateful to Dr. Olga Korosteleva at California State University, Long Beach, for her guidance on this paper, particularly regarding survival analysis methodology. Her mentorship has been invaluable to this project.

References

Alexander, Mariam, So Yeon Kim, and Haiying Cheng (Dec. 2020). "Update 2020: Management of Non-Small Cell Lung Cancer". In: *Lung* 198 (6), pp. 897–907. ISSN: 14321750. DOI: 10.1007/s00408-020-00407-5.

Azam, Faisal et al. (Sept. 2019). "Performance Status Assessment by Using ECOG (Eastern Cooperative Oncology Group) Score for Cancer Patients by Oncology Healthcare Professionals". In: *Case Reports in Oncology* 12 (3), pp. 728–736. ISSN: 16626575. DOI: 10.1159/000503095.

Braghetto, Anna et al. (Dec. 2022). "Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset". In: *Scientific Reports* 12 (1). ISSN: 20452322. DOI: 10.1038/s41598-022-18085-z.

Cerami, Ethan et al. (May 2012). "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data". In: *Cancer Discovery* 2 (5), pp. 401–404. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-12-0095.

Chevallier, Mathieu et al. (Apr. 2021). "Oncogenic driver mutations in non-small cell lung cancer: Past, present and future". In: *World Journal of Clinical Oncology* 12 (4), pp. 217–237. ISSN: 2218-4333. DOI: 10.5306/wjco.v12.i4.217.

Cramer-van der Welle, Christine M. et al. (Dec. 2021). "Real-world outcomes versus clinical trial results of immunotherapy in stage IV non-small cell lung cancer (NSCLC) in the Netherlands". In: *Scientific Reports* 11 (1). ISSN: 20452322. DOI: 10.1038/s41598-021-85696-3.

Davidson-Pilon, Cameron (2019). "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40, p. 1317. DOI: 10.21105/joss.01317. URL: https://doi.org/10.21105/joss.01317.

Ettinger, David S. et al. (May 2022). "Non-Small Cell Lung Cancer, Version 3.2022". In: *JNCCN Journal of the National Comprehensive Cancer Network* 20 (5), pp. 497–530. ISSN: 15401413. DOI: 10.6004/jnccn.2022.0025.

Germer, Sebastian et al. (Nov. 2024). "Survival analysis for lung cancer patients: A comparison of Cox regression and machine learning models". In: *International Journal of Medical Informatics* 191. ISSN: 18728243. DOI: 10.1016/j.ijmedinf.2024.105607.

Hardtstock, Fränce et al. (Mar. 2020). "Real-world treatment and survival of patients with advanced non-small cell lung Cancer: A German retrospective data analysis". In: *BMC Cancer* 20 (1). ISSN: 14712407. DOI: 10.1186/s12885-020-06738-z.

Ishwaran, Hemant et al. (Sept. 2008). "Random survival forests". In: *Annals of Applied Statistics* 2 (3), pp. 841–860. ISSN: 19326157. DOI: 10.1214/08-AOAS169.

Li, Ying et al. (Dec. 2022). "Machine learning models for identifying predictors of clinical outcomes with first-line immune checkpoint inhibitor therapy in advanced non-small cell lung cancer". In: *Scientific Reports* 12 (1). ISSN: 20452322. DOI: 10.1038/s41598-022-20061-6.

Malhotra, Jyoti et al. (June 2022). "Clinical outcomes and immune phenotypes associated with STK11 co-occurring mutations in non-small cell lung cancer". In: *Journal of Thoracic Disease* 14 (6), pp. 1772–1783. ISSN: 20776624. DOI: 10.21037/jtd-21-1377.

Mok, Tony S. et al. (Feb. 2017). "Osimertinib or Platinum–Pemetrexed in EGFR T790M–Positive Lung Cancer". In: *New England Journal of Medicine* 376 (7), pp. 629–640. ISSN: 0028-4793. DOI: 10.1056/nejmoa1612674.

Pölsterl, Sebastian (2020). "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn". In: *Journal of Machine Learning Research* 21.212, pp. 1–6. URL: http://jmlr.org/papers/v21/20-729.html.

Reck, Martin et al. (2016). "Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer". In: *New England Journal of Medicine* 375.19, pp. 1823–1833. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1606774.

Siegel, Rebecca L., Angela N. Giaquinto, and Ahmedin Jemal (Jan. 2024). "Cancer statistics, 2024". In: *CA: A Cancer Journal for Clinicians* 74 (1), pp. 12–49. ISSN: 0007-9235. DOI: 10.3322/caac.21820.

Soria, Jean-Charles et al. (Jan. 2018). "Osimertinib in Untreated EGFR -Mutated Advanced Non–Small-Cell Lung Cancer". In: *New England Journal of Medicine* 378 (2), pp. 113–125. ISSN: 0028-4793. DOI: 10.1056/nejmoa1713137.

U.S. Food and Drug Administration (2022). *FDA grants accelerated approval to fam-trastuzumab deruxtecan-nxki for HER2-mutant non-small cell lung cancer*. U.S. Food and Drug Administration. URL: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-accelerated-approval-fam-trastuzumab-deruxtecan-nxki-her2-mutant-non-small-cell-lung (visited on 10/22/2024).

Vanguri, Rami S. et al. (Oct. 2022). "Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer". In: *Nature Cancer* 3 (10), pp. 1151–1164. ISSN: 26621347. DOI: 10.1038/s43018-022-00416-8.

Yang, Tao et al. (2021). "Prognostic value of derived neutrophil-to-lymphocyte ratio (dNLR) in patients with non-small cell lung cancer receiving immune checkpoint inhibitors: a meta-analysis". In: *BMJ Open* 11.e049123, pp. 1–8. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2021-049123.

Yu, Hui et al. (2016). "PD-L1 Expression in Lung Cancer". In: *Journal of Thoracic Oncology* 11.7, pp. 964–975. ISSN: 1556-0864. DOI: 10.1016/j.jtho.2016.04.014.

Zhang, Chuan long et al. (2023). "Research progress and value of albumin-related inflammatory markers in the prognosis of non-small cell lung cancer: a review of clinical evidence". In: *Annals of Medicine*. DOI: https://doi.org/10.1080/07853890.2023.2192047.

Zhao, Wenhua et al. (Aug. 2021). "Impact of Smoking History on Response to Immunotherapy in Non-Small-Cell Lung Cancer: A Systematic Review and Meta-Analysis". In: *Frontiers in Oncology* 11. ISSN: 2234943X. DOI: 10.3389/fonc.2021.703143.