**AI-Guided Chemotherapy Optimization in Lung Cancer Using Genomic and Survival Data**

Moon, Phanny, Jenny, Minho, Liz

**Abstract**

The research aims to integrate various bioinformatics methodologies to develop prognostic genomic prediction markers for lung cancer patients. These predictive markers play a pivotal role in optimizing chemotherapy efficacy while minimizing adverse effects. This approach seeks to replace conventional one-size-fits-all medications with customized treatments tailored to individual patient needs. High-dimensional meta-database are constructed utilizing genomic datasets sourced from diverse independent clinical trials available in the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) data repository. To address population heterogeneity, new algorithms are considered. Beginning with univariate Cox regression for gene prescreening, ensemble methods incorporating regularized Cox Proportional Hazards (PH) models, Random Survival Forests, and Deep Survival Networks (DSNs) are utilized. These advanced computational algorithms aim to support complex clinical decision-making in cancer treatment. The anticipated outcome of this research is to significantly enhance cancer treatment survival outcomes by improving treatment efficacy and reducing toxicity risks in a specific patient population. Moreover, it offers invaluable support for navigating the complexities of clinical decision-making, thereby contributing to the advancement of personalized medicine.

1. **Introduction**

In 2018, global estimate from Global Cancer Observatory (GLOBOCAN) indicated that there were 2,094,000 new cases of lung cancer diagnosed worldwide, establishing lung cancer as the most common cancer globally (Bray et al., 2018). In the United States, according to the National Center Institute’s Surveillance, Epidemiology, and End Results (SEER) program, approximately 229,000 new cases of lung cancer were reported in 2020, representing 12.7% of all cancer diagnoses in the country (Thandra et al., 2021). Treatment options for lung cancer encompass surgery, radiation therapy, chemotherapy, and targeted therapy. In cases of Stage I and II resectable non-small cell lung cancer (NSCLC), surgery is a viable option. However, for Stage II resectable NSCLC, surgery followed by chemotherapy (adjuvant) is recommended (Lemjabbar-Alaoui et al., 2015). Nonetheless, it remains uncertain whether every individual in this category will benefit equally from chemotherapy. Physicians often explore a treatment option to enhance patient survival rates and improve their quality of life.

Despite undergoing curative resection, a significant proportion of patients with NSCLC, ranging from 30% to 55%, experience disease recurrence and succumb to their condition (Uramoto and Tanaka, 2014). Therefore, individuals facing a less favorable prognosis may benefit from adjuvant chemotherapy (ACT) (Zhu et al., 2010). Recent clinical studies have unveiled a survival advantage ranging from 4% to 15% for individuals with surgically removed Stages IB to IIIA when ACT is administered (Pirker and Filipits 2019). However, given the intrinsic toxicity of chemotherapy, physicians are responsible for cautiously identifying patients whose likelihood of benefiting from ACT justifies the associated risks.

Genomics plays a critical role in guiding treatment recommendations for lung cancer patients by providing insights into the specific genetic characteristics of tumors. This precision medicine approach allows for customized treatment strategies based on individual genomic profiles, leading to more effective therapies with fewer adverse effects. By identifying genomic markers associated with treatment response, clinicians can predict patient outcomes and personalized treatment plans accordingly. Additionally, genomic analysis helps identify mutations driving cancer growth, enabling the selection of targeted therapies that directly inhibit tumor progression. As a result, unveiling genomic markers associated with treatment for lung cancer patients is essential for advancing personalized medicine and improving treatment outcomes by optimizing treatment decisions based on individual genomic profiles.

In the treatment of NSCLC, identifying patient subgroups that respond optimally to specific therapies is crucial for enhancing treatment efficacy. The integration of artificial intelligence (AI) into clinical settings offers transformative potential for achieving this goal. By utilizing AI-driven techniques, researchers and clinicians can conduct detailed subgroup analyses that consider a wide array of factors, including clinical characteristics, genomic data, and genetic predispositions. This approach enables the identification of unique patterns and predictors of treatment success for the customization of therapy to individual patient profiles. AI tools can analyze vast datasets to uncover subtle interactions between treatment options and patient-specific factors, which often elude traditional analytical methods. Such precision allows for the adaptation of treatment plans that are finely tuned to maximize therapeutic benefits and minimize adverse effects. Ultimately, the application of AI in this context not only promises to improve the survival and quality of life for NSCLC patients but also represents a significant step forward in the evolution of precision medicine and personalized treatment strategies.

Moon et al. (2017) focused on identifying which patients might benefit from ACT based on prognostic and predictive gene signatures for ACT in early-stage NSCLC patients. The goal was to use gene expression profiling to develop a statistical decision-making algorithm for subgrouping patients who might be likely to benefit from ACT and those who might experience unnecessary toxicity, potentially allowing for more effective personalized treatment plans. The lasso method in Cox PH regression was used to identify significant genes associated with patient survival. A Random Forests classification model was applied to predict a new patient’s prognosis and the likelihood of survival benefit from ACT, using a set of genomic markers. The study aimed to improve the precision of ACT treatment decisions in order to optimize patient outcomes by minimizing unnecessary chemotherapy exposure and enhancing survival outcomes for those predicted to benefit.

Moon et al. (2020) initiated to develop a statistical decision support tool to identify risk factors and likelihood of benefit from ACT for early-stage lung cancer patients. Using genome-wide microarray data, the study identified genomic markers to predict treatment benefits and differentiated between patients who should receive chemotherapy alongside surgery and those who should undergo surgery alone. In this study, Accelerated Failure Time (AFT) models were used to predict the probability of benefiting from chemotherapy based on genomic markers. The comparison of treatment effects between a chemotherapy group and a surgery only group (OBS) revealed that patients who followed the predicted recommendation had significantly longer survival times (p < .001). Tree-based ensemble algorithms and Cox regression combined with the elastic net were applied to analyze gene expression data and to identify relevant genomic markers for better model performance. Overall, the study demonstrated how personalized treatment based on genetic profiles could optimize therapeutic strategies for lung cancer patients, potentially improving prognosis and survival outcomes.

Our recent study (Moon et al., 2021) continue to aim for identifying predictive genomic biomarkers that can distinguish subgroups of early-stage lung cancer patients likely to benefit from ACT. Utilizing a modified-covariate regularized Cox regression model with a lasso penalty on the JBR.10 dataset, the analysis revealed that patients following the predicted treatment based on genomic markers exhibited significantly higher survival benefits than those who did not (p < .001), emphasizing the potential of genomic markers in optimizing treatment decisions and reducing unnecessary costs and toxicities associated with ACT. The study identified specific genes like CDC42, ETV5 and FAM164A as potential predictive markers for treatment response, with ACT showing a more favorable prognosis compared to surgical resection alone (OBS) in advance cases, further highlighting the importance of the predictive model in guiding personalized treatment recommendations and improving survival outcomes for NSCLC patients. For robust results, further testing on new datasets and with more samples is essential to verify and refine the model's precision.

Our study sets the stage for a significant advancement in the field of clinical decision support for lung cancer treatment. By introducing innovative methodologies including a bagging approach based on the penalized Cox PH models, the Random Survival Forest (RSF) algorithm, and a deep learning survival network (DeepSurv), our goal is to enhance the robustness and accuracy of our predicted recommendation system. These advanced machine learning algorithms not only enable more precise predictions of treatment outcomes but also conforms recommendations to individual patient profiles. It eventually leads to optimize therapeutic strategies. Ultimately, this methodology promises to substantially improve clinical outcomes by facilitating more informed, personalized treatment decisions for patients with NSCLC.

1. **Methodology**

**2.1 Data Description**

The research to identify reliable gene signatures for lung cancer has been challenged by the heterogeneity and complexity of genomic data. Traditional approaches that rely on single datasets often fall short due to the lack of reproducibility and consistency across different studies. To address these issues, we have been implementing a meta-database approach in our research. This method aggregates and synthesizes data across multiple genomic datasets, which is crucial for overcoming the inherent variability and rare similarities found in individual studies. By using a meta-database, we aim to enhance the robustness and generalizability of our findings to provide a more accurate and comprehensive understanding of the specific roles that genes play in lung cancer pathology. This approach not only addresses the challenges posed by data diversity but also significantly enriches our analytical capacity to discern the most consequential gene signatures to facilitate more targeted and effective therapeutic interventions.

This research combined the following two datasets that are available at NCBI GEO data repository. First dataset GSE37745 (Botling et al., 2012) contains gene expression data for lung cancer patients from 1995 to 2005. This study focused on analyzing tissue samples from 196 patients with NSCLC to identify gene expression differences that could be linked to various clinical outcomes and characteristics of the disease. The dataset includes their tissue samples from both tumor and non-tumor using Affymetrix Human Genome U133 Plus 2.0 Array for comparative studies to understand cancer-specific genetic expressions better. The dataset also includes clinical data obtained from the area lung cancer registry. Patients had long-term follow-ups, which allowed clinicians to acquire additional data for the data set. In total, 71 patients had OBS, and 29 patients had ACT. The therapy received by the other 96 patients is unknown. As a result, only 100 patients with treatment information were used in this research.

The dataset, GSE29013 (Xie et al., 2011), details the gene expression profiles derived from 55 patients. The prepared samples were analyzed using the sophisticated Affymetrix U133 Plus 2.0 arrays, which are renowned for their comprehensive gene coverage and sensitivity. The study cohort comprised two distinct treatment groups: 21 of the patients received OBS treatment, typically involving regular monitoring without active intervention, while the remaining 34 patients were administered ACT. This division allowed the researchers to compare gene expression differences and potential biomarkers associated with treatment response.

**2.2 Model Design**

The datasets with 155 patients were split into training and testing sets, with 80% of the data used for training and the remaining 20% for test set. Specifically, for GSE29013, the training set consisted of 44 patients and the test set included 11 patients. For GSE37745, the numbers were 80 patients for training and 20 for test set. Compositionally, GSE29013 included 34 patients who received ACT and 21 who were OBS. GSE37745 had a different distribution with 29 ACT patients and 71 OBS patients. Table 1 presents a summary of the patient demographics from the combined datasets (GSE37745 and GSE29013) for both the learning and test data sets.

|  |  |  |
| --- | --- | --- |
| **TABLE 1.** Demographics of Combined Learning and Testing Data Sets | | |
|  | Learning Set (n = 124) | Testing Set (n = 31) |
| **Treatment Received** |
| Adjuvant Chemotherapy (ACT) | 50 | 13 |
| Observation (OBS) | 74 | 18 |
| **Age** |  |  |
| Less than 65 | 49 | 17 |
| Older than or equal to 65 | 75 | 14 |
| **Stage of Disease** |  |  |
| I | 74 | 15 |
| II | 25 | 8 |
| III | 24 | 8 |
| IV | 1 | 0 |

Each dataset was composed of 54,675 probe sets, which necessitated a meticulous screening process to identify which probe sets provided relevant information for predicting patient survival. This step was crucial to ensure the inclusion of valuable data while excluding elements that could introduce noise into the survival analysis. This approach aimed to refine the models and enhance their predictive accuracy regarding patient outcomes.

To streamline the dataset, the initial step involved performing Leave-One-Out Cross-Validation (LOOCV) alongside univariate analysis with Cox PH models to reduce variable count. Probe sets were selected based on a 5% significance level. Each probe received a variable importance score reflecting the frequency of its significance. The learning set was employed to implement each of the methods described below, using both fewer genomic variables and key clinical covariates such as age, sex and clinical stage.

**2.2.1 Regularized Cox Proportional Hazards Model**

The Cox proportional hazards (PH) model (Cox, 1972) is a widely used statistical technique in survival analysis. It is designed to examine the effect of several variables on the time until a specific event, such as failure or death, occurs. This model is particularly useful because it allows for the analysis for the simultaneous effect of several risk factors on survival. The core assumption of the Cox model is the proportionality of hazards, meaning the hazard ratios for any two individuals remain constant over time, even if their baseline hazard functions may differ.

In medical research, the Cox model enables researchers to assess the relative impact of variables such as treatment regimens, demographic factors, and biomarkers on the likelihood of an event like death or disease recurrence. Results from a Cox model are typically expressed in terms of hazard ratios (HR), which quantifies how a particular variable influences the risk of the event relative to the baseline. Hazard ratios provide an interpretable measure of relative risk.

The survival data is provided in the form where is the observed survival or censoring time, , indicates the censor status (1 for event, 0 for censored), and represents a covariate vector for the 𝑖-th individual, representing patient characteristics or treatment options. The Cox PH model is given by:

|  |  |
| --- | --- |
|  |  |

where is the baseline hazard rate when all covariates are zero, and is the regression coefficient of each covariate . The regression coefficients are estimated by maximizing the partial likelihood:

where represents the set of individuals at risk at time This partial likelihood approach allows for estimating the regression coefficients without needing to specify the baseline hazard function. Furthermore, the hazard ratio (HR) associated with treatment options (e.g., OBS) and (e.g., ACT) provides a measure of the relative risk:

|  |  |
| --- | --- |
|  |  |

where is the estimated model coefficients, and the exponent represents the difference in risk based on their respective treatments.

With genomic data, where the number of predictors greatly exceeds the number of observations, the Cox regression models often struggle with overfitting and inflated variance, leading to unreliable model predictions. To address these challenges, it becomes crucial to select the most informative variables while simultaneously regularizing the model to control complexity. Regularized Cox regression with an elastic net penalty (Zou and Hastie, 2005) offers a robust solution by combining the strengths of both (lasso) and (ridge) regularization techniques. This approach not only enhances variable selection by shrinking coefficients of less relevant predictors to zero but also retains stability in the presence of multicollinearity. This ensures a balance between interpretability and predictive performance. Thus, elastic net regularization provides a powerful framework for building reliable survival models in the context of high-dimensional survival data.

The Cox PH model, defined by its partial log-likelihood, is extended with an elastic net penalty that is expressed as:

where is the tuning parameter controlling the overall strength of regularization, is the mixing parameter balancing the (lasso, ) and (ridge, ) penalties, and are the Cox regression coefficients. The objective function for elastic net regularized Cox regression can be described as:

where is the penalized partial log-likelihood, is the linear predictor for the -th individual, is the risk set of individuals at time , is an event indicator (1 for event and 0 for censored) for the -th individual.

The hyperparameters and were optimized using a systematic approach during the training phase. A grid search was performed over values ranging from 0 to 1 in increments of .01, resulting in 101 candidate values. For each value of the corresponding optimal was determined using LOOCV. This produced 101 optimal values, one for each . The pair of and that achieved the best model performance was selected as the final hyperparameter combination.

We developed the first model using a bagging approach combined with regularized Cox regression. Bagging (Breiman, 1996) was employed to enhance the model’s stability and predictive performance by reducing variance. Each bootstrap sample was generated from the the training data with replacement to ensure a diverse representation of the data. For each bootstrap sample, a regularized Cox PH model was built. The predictions from individual models were aggregated to create the final ensemble model. Aggregation was performed by averaging the predicted risk scores across all bootstrap models to obtain a robust estimate of the survival risk for each patient.

Patient treatment recommendation is based on the predicted hazard ratio (HR), which quantifies the relative risk between two treatments options. Specifically, when HR > 1, hazard associated with (e.g., OBS) exceeds that of (e.g., ACT), favoring a recommendation for ACT. Conversely, if HR < 1, the hazard under is lower, supporting a recommendation for OBS. This approach provides a systematic and evidence-based framework for individualized treatment recommendations, aligning clinical decisions with patient-specific genomic and clinical risk profiles.

The model was implemented using the glmnet package (Friedman, et al., 2010) in R for elastic net regularization, and the bagging process was conducted using R.

**2.2.2 Random Survival Forests Model**

Random Survival Forests (RSF) (Ishwaran et al., 2008) is a nonparametric ensemble learning method specifically designed for survival analysis, extending the principles of random forests (Breiman, 2001) to handle time-to-event survival data. RSF builds upon bootstrap aggregation (bagging) and recursive partitioning to build a collection of survival trees using a random subset of predictors. Unlike traditional Cox PH models, RSF can handle more complicated patterns, like curvy relationships and how different factors interact, without needing strict rules like the proportional hazards assumption.

For a dataset with observations, a bootstrap sample is generated by randomly selecting observations with replacement. Let be the original dataset, where is the covariate for individual , and are the observed time and censoring indicators for individual , respectively. For each tree , a bootstrap sample is drawn from the training dataset with replacement, where Here, observations not included in the bootstrap sample are referred to as out-of-bag (OOB) samples. OOB samples are not used in the training of a particular tree. Each tree is grown using recursive partitioning on . At each node, a random subset of predictors is selected, and possible candidate splits are evaluated to maximize survival differences between child nodes. The log-rank statistic (Mantel, 1966) is used as a splitting criterion such that nodes are partitioned to best separate survival outcomes. The tree construction stops when a predefined minimum terminal node size is reached or when no further improvement in survival separation can be achieved.

The goal is to model the relationship between the covariates and survival outcomes by estimating survival probabilities or cumulative hazard functions (CHFs) for each patient. It aims to reduce variance and prevent overfitting. The result is a set of survival trees, each capturing different aspects of the data structure due to the randomness introduced by bagging.

After trees are grown, their predictions are aggregated to produce a final RSF model. For an individual patient , the RSF estimates the cumulative hazard function (CHF) for each treatment (e.g., for OBS and for ACT) and takes an average of the CHFs from all trees:

where is the CHF predicted by tree for treatment , and is the total number of trees in the forest.

For an individual , the RSF model can provide survival probabilities at any given time . These probabilities are derived by aggregating the survival probabilities across all trees in the forest. Using the aggregated CHFs, the survival probability for patient under treatment at time , is obtained as:

where is the CHF at time for treatment .

RSF’s ability to handle large number of predictor variables while maintaining predictive accuracy makes it an ideal choice as the second model in this paper for individualized clinical decision-making in the treatment of non-small cell lung cancer patients. Out of the key features of RSF is its ability to estimate variable importance, which indicates the contribution of each predictor to the model. Variable importance is assessed by permuting the values of a specific predictor in the OOB samples and calculating the decrease in prediction performance. Predictors with higher importance scores are more influential in determining survival outcomes.

For each patient, the survival probabilities and are compared across a range of time points. If the OBS treatment is recommended. On the other hand, if the ACT treatment is recommended.

The RSF model was implemented using the randomForestSRC package in R. All hyperparameters, including the number of trees, the random split points, the number of variables randomly selected at each split in a node, and the minimum terminal node size, were optimized through LOOCV to ensure robust model performance.

**2.2.3 Deep Learning Survival Model**

DeepSurv (Katzman et al., 2018) is a deep learning survival model that combines the principles of Cox regression with deep neural networks. By learning complex nonlinear relationships and interactions among predictor variables, DeepSurv extends the capabilities of traditional Cox PH models for personalized survival analysis and treatment recommendations. The goal is to model the hazard function for a patient as:

where is the baseline hazard, and is log hazard ratio predicted by the neural network with weight parameters .

The DeepSurv model utilizes a feedforward neural network (FNN) to predict the log hazard ratio, , for each patient. The network architecture consists of an input layer, hidden layers and output layer. Some dropout layers and a regularization are employed to prevent overfitting and improve generalization. The input layer takes the patient-specific covariate vector as input. One or more fully connected hidden layers with nonlinear activation functions are used to capture complex interactions and nonlinear relationships between predictors. A single output node, without activation function, predicts the log hazard ratio. Figure 1 shows an example of DeepSurv with 32 input units, 2 hidden layers with 8 and 4 units, respectively, and output layer.

|  |
| --- |
| A diagram of a network  Description automatically generated |
| Figure 1. An example of DeepSurv architecture with two hidden layers. |

The loss function in DeepSurv can be derived from the negative partial likelihood of the Cox regression model as equation (1) in Section 2.2.1, which is adapted to train the deep neural network. The partial likelihood focuses on the relative risk among individuals without modeling the baseline hazard function. Taking the logarithm gives the partial log-likelihood

where is a set of individuals who experienced the event and is individuals still at risk at time . To define a loss function for optimization, we minimize the negative partial log-likelihood, .

To improve the generalization and stability of the model, elastic net regularization is incorporated into the loss function, where where controls the sparsity of the model by regularization, and penalizes large weights to prevent overfitting by regularization. Thus, the final loss function would be:

The loss function is minimized using an optimization algorithm such as stochastic gradient descent (SGD) algorithm. Gradients of with respect to is used to update the weights via backpropagation to propagate errors through the network as follows:

where is the learning rate. Model is trained on the training dataset, and hyperparameters (e.g., learning rate, regularization terms, batch normal, hidden layers, dropout rates, and activation) are tuned based on performance of the validation set.

Bayesian hyperparameter optimization (BHO) (Sneok et al., 2012) was utilized to initiate the parameter search process. BHO is an efficient approach as it provides a scientific and efficient method for hyperparameter tuning. It handles high-dimensional spaces effectively and is also widely adopted in machine learning and deep learning hyperparameter tuning. While grid search or random search evaluate hyperparameters in a brute-force or random manner, BHO uses probabilistic models to efficiently explore the hyperparameter space. It is particularly useful when evaluating each combination of hyperparameters is computationally expensive.

Jepkoech et al. (2021) emphasized that improper selection of learning rates can significantly impact model training. High learning rates may result in unpredictable model behavior and hinder training accuracy, while excessively low learning rates can slow training progress due to minimal weight updates. Thus, determining an optimal learning rate is critical and often achieved by testing and evaluating different values.

The structure of a neural network is primarily defined by its hidden layers, including both the number of layers and the number of nodes in each layer. Dropout is a regularization technique used during training to mitigate overfitting by randomly removing certain nodes, effectively reducing redundant or excessive connections between neurons.

Batch normalization (BN) is another key technique that standardizes activations across intermediate layers of the network, improving accuracy and accelerating training (Bjorck et al., 2019). Additionally, scaled exponential linear units (SELU) serve as activation functions that enable automatic normalization of network activations, enhancing training stability and improving convergence (Klambauer et al., 2017).The model was implemented in Python using deep\_surv package (Katzman et al., 2018).

Using this model, treatment recommendations are based on the logarithmic transformation of the hazard ratio defined in Equation (2). For each patient, the algorithm estimates the risk level associated with a specific treatment. Assuming a common baseline level of risk, the relative risks of different treatments are compared by calculating the logarithm of the hazard ratio. This calculation yields the "recommender function," denoted as , which quantifies the difference in log hazards between two treatment options for an individual:

where and are the estimated log hazards for treatments (e.g., OBS)and (e.g., ACT), respectively.

To generate the recommendation, the algorithm evaluates the patient's data under both treatment groups(OBS)and (ACT) and compares the results. A positive value of indicates that treatment (OBS) is associated with a higher risk of an adverse outcome compared to treatment (ACT), leading to a recommendation for ACT. Conversely, a negative value of suggests that treatment (OBS) is more effective, as it is associated with a lower risk of an adverse outcome than treatment (ACT).

**2.2.4 Performance measure: Concordance Index**

The Concordance index (C-index) was used to quantify the model’s performance (Harrell, et al., 1982). The C-index is a widely used performance matric in survival analysis that quantifies the discriminative ability of a predictive model. It evaluates the extent to which a model correctly ranks survival times, determining whether patients with higher predicted risks experience events (e.g., death) earlier than those with lower predictive risks. The C-index provides an objective assessment of prognostic models used for individualized treatment recommendations (Uno et al., 2011).

The C-index is defined as:

Here, represents the set of all permissible (comparable) patient pairs , where (i.e., patient experiences an event before patient ). The denotes the predicted risk score or hazard function derived from the survival model. The is the indicator function, which takes a value of 1 if the predictive risk score satisfies , and 0 otherwise. The is the total number of comparable patient pairs.

A C-index of .5 indicates that the model’s predictions are equivalent to random choice, while a C-index of 1.0 represents perfect concordance, meaning that the model correctly ranks all patient survival times (Harrell et al., 1996). In practice, a C-index above .7 is generally considered indicative of a model with good discriminative ability (Pencina and D’Agostino, 2004), even though .8 and above are desirable.

For survival data with right-censored observations, Harrell’s C-index is commonly used, as it accounts for censored patients by excluding non-informative pairs where the event time is not observed (Steyerberg et al., 2018). This adaptation ensures robustness when evaluating survival models on clinical datasets where censoring is prevalent.

In the context of personalized medicine for NSCLC, the C-index serves as a crucial performance metric, ensuring that survival predictions used in clinical decision making are reliable and can guide evidence-based, individualized treatment strategies (Ishwaran et al., 2008).

1. **Results**

To assess the predictive performance of our survival models, we computed the C-index on the test dataset for three proposed models: a bagging with regularized Cox regression, RSF, and DeepSurv. This section evaluates their effectiveness in optimizing patient-specific survival predictions and treatment recommendations.

For feature screening, we applied a treatment interaction Cox regression model to identify relevant predictors of patient treatment outcomes. Using LOOCV on the training dataset (described in Table 1), we initially excluded probe sets with out of the 54,675 probe sets from further analysis After screening, a refined subset of 1,834 probe sets and 4 clinical and demographic variables—age, sex, treatment, and stage—was retained. These screened variables were consistently used across all three survival models to ensure a generalized comparison framework.

In the bagging with regularized Cox model, we generated 200 bootstrap resamples during the training phase. Based on the estimated risk scores, treatment recommendations were assigned. Within the training dataset, the model predicted 82 patients to the OBS category and 42 patients to the ACT category. To evaluate the effectiveness of these recommendations, we analyzed patient survival curves, comparing those who adhered to the model’s treatment recommendation with those who did not. Among patients in the training dataset, 112 followed the model recommendation (either ACT or OBS), while 12 did not. Similarly, in the test dataset, the model classified 29 patients into the OBS group and 2 patients into the ACT group. Among them, 16 patients actually followed the model’s recommendation, while 15 did not.

The discriminative ability of the bagging with regularized Cox model resulted in a C-Index of .996 on the training dataset and .709 on the test dataset, indicating strong predictive performance. The survival probabilities for patients who followed the model’s treatment recommendation (solid line) compared to those who did not (dashed line) are illustrated in Figure 2. Patients who adhered to the model’s recommendation maintained a higher survival probability throughout approximately 11 years. However, both groups experience deaths toward the 12-16 year range. The median survival time is estimated at 13.4 years. While there is a visible difference in survival probability up until around 11-year mark, the log-rank test comparing two groups survival probabilities resulted in 0, indicating not enough evidence for statistically significant difference in survival distributions between those who followed versus did not follow the model recommendation.

Despite the higher survival probability in the followed group, the lack of statistical significance suggests that several factors may be at play. One possibility is the small sample size, which may have limited the statistical power needed to detect a meaningful difference in survival outcomes. We noted that only 2 patients remained at risk in the followed group by year 12. Additionally, the model’s recommendations may not provide strong enough differentiation between patient survival, potentially indicating the need for further refined model.

|  |
| --- |
| A graph with numbers and a line  Description automatically generated  0 |
| **FIGURE 2.** Survival curves of patients following versus not following model recommendations using Cox proportional hazards with Log-rank test. |

As a further refined model, RSF was employed with the same training and test data in Table 1. Hyper parameters are tuned in training phase via 10-fold cross-validation for enhancing model performance and ensuring accurate survival predictions. Key hyperparameters were the total trees in the forest (ntree), the number of variables for splitting that specifies the subset of features evaluated at each split node (mtry), and node size to determine the minimum sample size required to split a node (nsplit) and to be a leaf node (nodesize), Splitting criterion to determine the best splits was determined by the log-rank test. For this hyperparameterization, we used R with randomForestSRCand caret packages. In this step, the following parameters were determined: mtry = 37, nodesize = 6, ntree = 1000, and nsplit = 10.

Based on RSF outcome for treatment recommendations, 83 patients were recommended to OBS category and 41 patients were recommended to ACT category in the training data. This resulted in 75 patients actually followed the model’s recommendation and 49 did not. Similarly, for the test dataset, 20 patients were recommended to the OBS group, whereas 11 patients were recommended to the ACT group. Thus, 17 patients actually followed the model’s recommendation, and 14 patients did not follow. For the model’s performance, the mean C-Index for the training set is .889, and the mean CI for the test set is .885.

Table 2 presents the variable importance scores from the RSF model, indicating the relative contribution of each predictor to the treatment-related survival prediction. The higher the importance score, the more influential the variables is in predicting survival outcomes. For example, TTR (transthyretin) has the highest variable importance score, suggesting a strong association with survival outcomes. This gene is known for a biomarker for neurodegenerative diseases and systemic amyloidosis. PREPL (prolyl endopeptidase-like is usually involved in neurological and metabolic pathways suggesting further exploration of its role in disease progression. MTURN is a neural progenitor differentiation regulator, playing role in cell development that may influence cancer progression through its impact on cell differentiation and growth pathways. The identified genes could serve as predictive biomarkers for survival outcomes, potentially influencing treatment strategies.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2**. Random Survival Forest Variable Importance Score | | | |
| Variable | Variable Importance Score | Gene  Symbol | Gene  Descriptions (SynGo Consortium, 2019) |
| 209660\_at | 0.008555592 | TTR | transthyretin |
| 212215\_at | 0.007373110 | PREPL | prolyl endopeptidase like |
| 227000\_at | 0.007348966 | MTURN | maturin, neural progenitor differentiation regulator homolog |
| 227200\_at | 0.006915272 | ETV3 | ETS variant transcription factor 3 |
| 218811\_at | 0.006702557 | ORAI2 | ORAI calcium release-activated calcium modulator 2 |
| 228886\_at | 0.006036600 | LRRC27 | leucine rich repeat containing 27 |
| 240184\_at | 0.006008361 | SYNPR-AS1 | SYNPR antisense RNA 1 |
| 218230\_at | 0.005832459 | ARFIP1 | ADP ribosylation factor interacting protein 1 |
| 225012\_at | 0.005657217 | HDLBP | high density lipoprotein binding protein |
| 205504\_at | 0.005625274 | BTK | Bruton tyrosine kinase |
| 208683\_at | 0.005619357 | CAPN2 | calpain 2 |
| 203126\_at | 0.005215127 | IMPA2 | inositol monophosphatase 2 |
| 225273\_at | 0.005053145 | WWC3 | WWC family member 3 |
| 207249\_s\_at | 0.004782021 | SLC28A2 | solute carrier family 28 member 2 |
| 206211\_at | 0.004511744 | SELE | selectin E |
| 229145\_at | 0.004417019 | ANAPC16 | anaphase promoting complex subunit 16 |
| 226146\_at | 0.004406515 | HEIH | hepatocellular carcinoma up-regulated EZH2-associated long non-coding RNA |
| 235352\_at | 0.004392154 | MR1 | major histocompatibility complex, class I-related |
| 234297\_at | 0.004381663 | RGS8 & SDHAP3 | regulator of G protein signaling 8 & SDHA pseudogene 3 |
| 224650\_at | 0.004321009 | MAL2 | mal, T cell differentiation protein 2 |
| 218693\_at | 0.004226230 | TSPAN15 | tetraspanin 15 |
| 218707\_at | 0.004063735 | ZNF444 | zinc finger protein 444 |
| 233167\_at | 0.003895603 | SELENOO | selenoprotein O |
| 209682\_at | 0.003893244 | CBLB | Cbl proto-oncogene B |
| 200667\_at | 0.003872056 | UBE2D3 | ubiquitin conjugating enzyme E2 D3 |
| 229970\_at | 0.003856160 | KBTBD7 | kelch repeat and BTB domain containing 7 |
| 219468\_s\_at | 0.003790877 | CUEDC1 | CUE domain containing 1 |
| 205448\_s\_at | 0.003734501 | MAP3K12 | mitogen-activated protein kinase kinase kinase 12 |
| 201236\_s\_at | 0.003712341 | BTG2 | BTG anti-proliferation factor 2 |
| 214623\_at | 0.003701552 | FBXW4P1 | F-box and WD repeat domain containing 4 pseudogene 1 |
| 221861\_at | 0.003696707 | ARRB1 | arrestin beta 1 |
| 241208\_at | 0.003691053 | PDLIM5 | PDZ and LIM domain 5 |

Figure 3 illustrates the difference in survival probabilities between two patient groups: those who followed the treatment recommendations provided by the RSF (solid line) and those who did not follow the model’s recommendations (dashed line). The group that followed the RSF model recommendations demonstrates a higher survival probability over the time compared to the non-followed group. The separation between the two survival curves suggests a potential benefit of adhering to the model’s guidance . The median survival time for the is 13.4 years. The not-followed group experiences a more rapid decline in survival probability, indicating a worse prognosis for those who did not follow the recommended treatment strategy.

|  |
| --- |
| A graph with numbers and lines  Description automatically generated |
| **FIGURE 3.** Survival curves of patients following versus not following model recommendations using Random Survival Forests with Log-rank test. |

The results from RSF suggest that following the RSF model’s recommendations is associated with improved survival outcomes. In this paper, further validation is conducted using a deep-learning survival network to confirm these results.

Our third model was DeepSurv. In training phase with train datasets, hyperparameters were decided as follows: learning rate = 0.3, dropout = .5, learning rate decay = 1.0, L2 regularization = 2.46, L1 regularization = 4.14, batch norm = True, two hidden layers of sizes = [100, 50], standardize = True, and activation = SELU. Thus we had 1838 input neurons, two hidden layers of size 100 and 50, and one output layer in the survival neural network.

DeepSurv recommended 1 patient to OBS category and 92 patients in ACT category in the training set. This resulted in 56 patients who actually followed the model’s recommendation (either ACT or OBS) and 37 who did not. Similarly, for the test set, 2 patients were recommended to the OBS group, whereas 29 patients were recommended to the ACT group. Among the test set, 16 patients actually followed the model’s recommendation, and 15 patients did not follow. The C-index for DeepSurv was .990 in the training data and .982 for the test data.

Figure 4 presents the survival curves generated using the DeepSurv model, comparing the survival probabilities of patients who followed (solid line) versus did not follow (dashed line) the model’s treatment recommendations. The survival curves for the Followed and Not-Followed groups show minimal separation. The median survival time was estimated to 13.4 years

“Followed” group (solid line) initially shows slightly better survival than the “Not-Followed” group (dashed line) but eventually converges near the end. Even though the trend is similar to RSF results, but the separation between the two groups is less pronounced than in the RSF model. The median survival time in this model is estimated to 10.9 years that is shorter than one from RSF model. The lack of significance may be because by time 10, very few patients remain in either group, which may affect the reliability of estimates at later time points.

|  |
| --- |
| A graph with numbers and lines  Description automatically generated  0 |
| **FIGURE 4.** Survival curves of patients following versus not following model recommendations using DeepSurv with Log-rank test. |

1. **Conclusion**

This study presents an advanced clinical decision support framework for optimizing adjuvant chemotherapy (ACT) treatment in non-small lung cancer (NSCLC) patients. By integrating high-dimensional genomic datasets with advanced machine learning algorithms such as regularized Cox regression models, Random Survival Forests (RSF), and Deep Survival Networks (DeepSurv), we have developed a robust and scalable predictive model for guiding personalized chemotherapy decisions. Our findings demonstrate that machine learning-driven survival analysis can accurately predict patient subgroups that are more likely to benefit from ACT, thereby minimizing unnecessary chemotherapy exposure while improving survival outcomes.

Through a systemic evaluation using multiple independent datasets from the National Center for Biotechnology Information (NCBI), we validated that our predictive models exhibit high discriminatory power (as measured by concordance index) and can significantly enhance the precision of treatment recommendations. The survival analysis highlights the clinical relevance of genomic biomarkers such as TTR, MTURN and ETV3, which emerged as key predictors of patient response to ACT. These biomarkers in Table 2 warrant further investigation targets for biomarker-driven therapeutic strategies in lung cancer treatment.

From a broader perspective, our study underscores the transformative potential of artificial intelligence (AI) in precision oncology. By utilizing AI-driven approaches, clinicians can better stratify NSCLC patients based on genomic risk factors, ensuring that chemotherapy interventions are more targeted and clinically justified. This advancement has profound implications for reducing chemotherapy-related toxicity, improving patient quality of life, and optimizing healthcare resource allocation. Future research should focus on expanding the model’s applicability to larger, more diverse patient cohorts and integrating additional multi-omics data to refine predictive accuracy further.

In conclusion, this study provides a scientifically rigorous and clinically relevant framework for data-driven chemotherapy decision-making in NSCLC. Our work paves the way for future advancements in personalized lung cancer therapy, demonstrating the potential of genomic-guided treatment selection to significantly improve survival rates and treatment efficacy in lung cancer patients.

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Author contribution(s)**

Hojin Moon and Lauren Tran had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

*Concept and design: Moon and Tran.*

*Acquisition, analysis or interpretation of data: All authors.*

*Drafting of manuscript: Moon and Tran.*

*Critical Review of the Manuscript for important intellectual content: Moon and Tran.*

*Statistical analyses and Visualization: All authors.*

*Administrative, technical, or material support: Moon*

*Supervision: Moon*

All authors have read and agreed to the published version of the manuscript.

**Acknowledgments**

Hojin Moon’s research was partially supported by the Research, Scholarship, and Creative Activity (RSCA) Award from CSULB.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**Declaration of conflicting interest**

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Availability of data and materials**

Data are available upon reasonable request for access to the datasets used in this study.

**Reference**

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
2. Thandra KC, Barsouk A, Saginala K, Aluru JS, Barsouk A. Epidemiology of lung cancer. *Contemp Oncol (Pozn)*. 2021;25(1):45-52.
3. Lemjabbar-Alaoui H, Hassan OU, Yang YW, Buchanan P. Lung cancer: Biology and treatment options. *Biochim Biophys Acta*. 2015;1856(2):189-210.
4. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res* 2014;3(4):242-249.
5. Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, Pennell N, Thomas RK, Naoki K, Ladd-Acosta C, Liu N, Pintilie M, Der S, Seymour L, Jurisica I, Shepherd FA, Tsao MS. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol*. 2010;28(29):4417-24.
6. Pirker R, Filipits M. Adjuvant Therapy in Patients With Completely Resected Non-small-cell Lung Cancer: Current Status and Perspectives. *Clin Lung Cancer*. 2019;20(1):1-6.
7. Moon H, Zhao Y, Pluta D, Ahn H. Subgroup analysis based on prognostic and predictive gene signatures for adjuvant chemotherapy in early-stage non-small-cell lung cancer patients. *J Biopharm Stat.* 2018;28(4):750-762.
8. Moon H, Chao T, Ahn H. Identification of Risk Factors and Likelihood of Benefit from Adjuvant Chemotherapy for Early Stage Lung Cancer Patients. *J Biopharm Stat.* 2020 May 3;30(3):430-444.
9. Moon H, Nguyen A, Lee E. Prognostic Genomic Predictive Biomarkers for Early-Stage Lung Cancer Patients. *The Open Biomarkers Journal*. 2021;11(1):69–78.
10. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological).* 1972;34(2):187–202.
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-20. doi:10.1111/j.1467-9868.2005.00503.x.
12. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324.
13. Breiman L. Bagging predictors. *Machine Learning*. 1996;24(2):123–40. doi:10.1007/BF00058655.
14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010;33(1):1–22. doi:10.18637/jss.v033.i01.
15. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-60. doi:10.1214/08-AOAS169.
16. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*. 1966;50(3):163–70.
17. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*. 2018;18(1):24. doi:10.1186/s12874-018-0482-1.
18. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. *Adv Neural Inf Process Syst*. 2012;25:2951–9.
19. Jepkoech J, Mugo DM, Kenduiywo BK, Too EC. The effect of adaptive learning rate on the accuracy of neural networks. *Int J Adv Comput Sci Appl*. 2021;12(8). Available from: <https://doi.org/10.14569/ijacsa.2021.0120885>
20. Bjorck J, Gomes C, Selman B, Weinberger K. Understanding Batch Normalization. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS’18). Red Hook, NY: Curran Associates Inc.: 2018:7705–16. Available from: <https://dl.acm.org/doi/pdf/10.5555/3327757.3327868>.
21. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, CA: Curran Associates Inc.; 2017:971-80. Available from: <https://proceedings.neurips.cc/paper/2017/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf>
22. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543–6. doi:10.1001/jama.1982.03320430047030.
23. Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistic for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105–1117. doi:[10.1002/sim.4154](https://doi.org/10.1002/sim.4154)
24. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–387. doi:[10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(sici)1097-0258(19960229)15:4%3C361::aid-sim168%3E3.0.co;2-4)
25. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. Stat Med. 2004;23(13):2109–2123. doi:[10.1002/sim.1802](https://doi.org/10.1002/sim.1802)
26. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and their evaluation in the presence of censoring. Stat Methods Med Res. 2018;27(9):2504–2525. doi:[10.1016/j.recesp.2011.04.017](https://doi.org/10.1016/j.recesp.2011.04.017)
27. SynGO Consortium. SynGO - ID Conversion Tool. SynGo.org; 2019. Available from: <https://www.syngoportal.org/convert>.