International Journal of Data Science and Analysis
2021; 7(3): 60-68
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20210703.12
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Developing Genomic Predictive Biomarkers for Survival Benefit from Adjuvant Chemotherapy in Early-Stage Lung Cancer Patients for Personalized Medicine

Hojin Moon[1, *], Evan Lee[2]

[1]Department of Mathematics and Statistics, California State University, Long Beach, California, the United States

[2]Yale University, New Haven, Connecticut, the United States

Email address:

hojin.moon@csulb.edu (H. Moon), evan.lee@yale.edu (E. Lee)

[*]Corresponding author

To cite this article:

Hojin Moon, Evan Lee. Developing Genomic Predictive Biomarkers for Survival Benefit from Adjuvant Chemotherapy in Early-Stage Lung Cancer Patients for Personalized Medicine. *International Journal of Data Science and Analysis*. Vol. 7, No. 3, 2021, pp. 60-68.
doi: 10.11648/j.ijdsa.20210703.12

Received: February 1, 2021; **Accepted**: February 8, 2021; **Published**: May 8, 2021

Abstract: Surgical resection only remains the standard choice for the treatment of early-stage non-small cell lung cancer (NSCLC) patients. Preliminary studies suggest that the application of adjuvant chemotherapy with surgery (ACT) is associated with a better prognosis for more severe NSCLC patients compared to those who only underwent surgical resection. However, at an individual level, not all patients may benefit from ACT. Given the well-known adverse effects and toxicity of ACT, finding the patients that are most likely to benefit from ACT is paramount. Thus, the purpose of this research is to utilize gene expression and clinical data from lung cancer patients to develop a statistical decision support algorithm to find predictive genomic biomarkers and identify subgroups of patients who benefit from ACT. Cox regression models are trained using a randomized controlled trial gene expression data from the National Center for Biotechnology Information (NCBI) utilizing explicit treatment interaction terms. To handle high dimensions inherent in gene expression data, a regularized Cox regression model with lasso penalty is applied to find the most significant interacting markers. Risk scores are estimated from the proposed model and are used to stratify patients into a high risk or low risk group respective to ACT treatment. After applying the model to an independent validation genomic data set, we show that patients who underwent the recommended treatment according to their risk group estimated by our proposed algorithm exhibit a slightly higher survival rate than those who do not.

Keywords: Decision Support Algorithm, Genomic Markers, Regularized Cox Model, Subgroup Analysis, Survival Analysis

1. Introduction

Lung cancer is not only among the most widely diagnosed types of cancer (only falling second to breast cancer, 2.21 million cases vs. 2.26 million cases in 2020), but it also leads in deaths with 1.80 million in 2020 [1]. The most common type of lung cancer is non-small cell lung cancer (NSCLC), which accounts for 84% of diagnoses. Although each stage of cancer has its general, standard treatment, it is of utmost importance to consider treatment decisions on a patient-by-patient basis in order to optimize patient survival rate. In recent years, much attention has been ascribed to studying genomic data and identifying genomic markers that help to predict how patients will respond to various treatment options. This information allows doctors to make more educated treatment decisions based on the individual, ultimately increasing overall patient outcome and maximizing the efficacy of the treatment.

Currently, the stage of a lung cancer patient is arguably the most important factor in deciding what treatment should be undertaken. Surgery only is generally recommended for patients with Stage I lung cancer, while adjuvant chemotherapy treatment after surgery (ACT) is generally recommended for patients with Stage II or III lung cancer [2]. There are several factors like tumor size for Stage I patients or comorbidities that may change this recommendation. However, a recommendation of chemotherapy for Stage I patients continues to be controversial. Although the value of adjuvant chemotherapy has become widely accepted as a result

of clinical research from the past few decades, adjuvant chemotherapy elicits devastating side effects on many patients. A vast majority of patients in the JBR. 10 study experienced side effects like neutropenia (88% of patients), fatigue (81%), nausea (80%), and neuropathy (48%). Similarly, patients in the ANITA study experienced neutropenia (92%) and grade 3-4 nausea/vomiting (27%) [3, 4]. The prevalence and severity of the side effects of chemotherapy warrant further research into the use of genomic markers—specifically, to predict how an individual will respond to adjuvant chemotherapy.

With the development of genetics research and the increasing number of publicly available data sets in recent years, researchers now have better access to the tools necessary to identify such genomic markers and improve treatment efficacy. These resources have brought about the rise of bioinformatics and computational biology. The development of these fields has made way for medical strides not only in terms of lung cancer but also in various other illnesses and diseases. Specifically, the use of biomarkers to improve patient outcomes has continued to draw significant interest in recent years. For instance, Ibrahim et al. [5] established that the identifying of biomarkers is expected to be critical in continuing to improve the care of heart failure patients. Although researchers have made some progress in the use of biomarkers, it is clear that there is still much progress to be made. While multiple biomarkers that are significantly correlated to cardiovascular risk have been identified, their role in improving risk prediction still remains limited [6]. In terms of NSCLC, some potential biomarkers have been identified in research, but the potential use of them in the management of lung cancer patients requires further exploration [7]. The promising future of genomic markers and their potential to create profound impacts in improving patient care make them increasingly important to investigate.

A goal of this study is to create a prognostic gene signature by identifying a set of treatment-related genomic biomarkers. Previous studies have had similar aims; For instance, He et al. [8] identified an 8-gene signature in 2019, and Zuo et al. [9] identified a 6-gene signature in 2019 for NSCLC. Similarly, Boutros et al. [10] selected and validated a 6-gene signature, and noted that there are hundreds of thousands of other verifiable NSCLC prognostic signatures. The existence of so many verifiable signatures due to differing statistical methods serves to show why there is often a lack of overlap from study to study. The aforementioned studies, for instance, have minimal overlap in their identified gene signatures.

Gene signatures themselves have many important applications. For cancer, gene signatures aid physicians in predicting recurrence and the speed at which cancerous cells can grow and spread in an individual's body. They may also be able to predict treatment efficacy so the optimal treatment plan can be utilized, in addition to helping diagnose disease and accurately determining patient prognosis. These applications make gene signatures critical for the improvement of overall patient care.

Moon et al. [11, 12] previously investigated the identification of a gene signature with the goal of recommending treatments that would lead to better survival outcomes. They built two separate models by splitting their training set into two smaller sets based on treatment. However, the splitting of the larger data set meant that the models were built on smaller training sets, which may result in the selection of non-optimal genomic markers. Their prediction algorithm also resulted in inconclusive recommendations for certain patients because the two model recommendations had not agreed with each other.

In this paper, we propose using a lasso-regularized Cox regression model with treatment interaction in order to estimate the risk of taking ACT, thereby minimizing the toxicity and improving patient survival. Specifically, if the risk of adjuvant chemotherapy is lower than the risk of surgery only, ACT is recommended, and vice versa.

In Section 2, the training data and a separate validation dataset are described and preprocessed. Methodologies used in this paper are illustrated in Section 3. Results are reported in Section 4. We conclude our findings in Section 5.

2. Data Description and Preparation

For the training set, we utilized a 442-patient data set [13]. The data set was considered to be the largest publicly available microarray data set with significant lung adenocarcinoma annotation. The data set utilized TN staging, which was inconsistent with the stages of I, II, and III used in the validation set. For this reason, the TN stages from the raw data of the training set were converted to stages I, II, and III using the American Joint Committee on Cancer's Lung Cancer Staging. pN0pT1 and pN0pT2 were converted to stage I; pN0pT3, pN1pT1, and pN1pT2 were converted to stage II; pN0pT4, pN1pT3, pN1pT4, pN2pT1, pN2pT2, pN2pT3, and pN2pT4 were converted to stage III. These conversions were are denoted in Table 1. After the removal of 56 patients due to missing data points, it was found that 244 were stage I patients, 77 were stage II patients, and 65 were stage III patients. The raw data can be downloaded using accession number GSE68465 on the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68465).

Analyzing the training data using the Kaplan-Meier estimator, we found that there was a statistically significant difference in the survival rates of the patients who underwent surgery only and the patients who underwent adjuvant chemotherapy (p=0.00019, Figure 1). Those who underwent surgery only had a median survival time of 6.31 years, while those who underwent adjuvant chemotherapy had a median survival time of 3.77 years. The data itself showed an adverse effect on ACT that did not prolong patient survival. This motivated our paper to develop an algorithm to select a subgroup of patients who could benefit from ACT.

We used a smaller, 133-patient data set for the validation set [14]. The data set was a randomly selected subset of the data from the JBR. 10 trial, which was originally composed of 482 patients [3]. Out of 133 patients, 62 underwent surgery only

and the remaining 71 underwent adjuvant chemotherapy. The raw data can be downloaded using accession number GSE14814 on the NCBI website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14814). Information for the two sets is displayed in Table 2.

Affymetrix microarrays have become a popular means of performing gene expression analysis as they are capable of simultaneously detecting the expression of thousands of genes. Because of this ability, they have become a critical tool in the field of genomics. Microarrays themselves are composed of microscopic spots imprinted on microscopic slides. Every microscopic spot on the slides corresponds to some known DNA gene or sequence. Genes on a GeneChip are represented by probe sets.

Raw microarray data must go through preprocessing before analysis. Typically, they first undergo background correction, followed by normalization, and finally summarization. Each step has a specific purpose; Background correction allows for the reduction of the effect of local artifacts and other noise. The process helps with the elimination of spatial heterogeneity or non-specific binding effects, but may cause issues like corrected intensities with negative value [15]. The normalization process helps to reduce bias and error that may result from how the data is acquired, such as differences in technology. Reducing such bias allows measurements from one array to be comparable to measurements from another array. The main purpose of summarization is to summarize probe readings in a singular number as a concise means of representing gene expression. These three steps of preprocessing were performed on the raw training and validation data using Bioconductor's "affy" package in R.

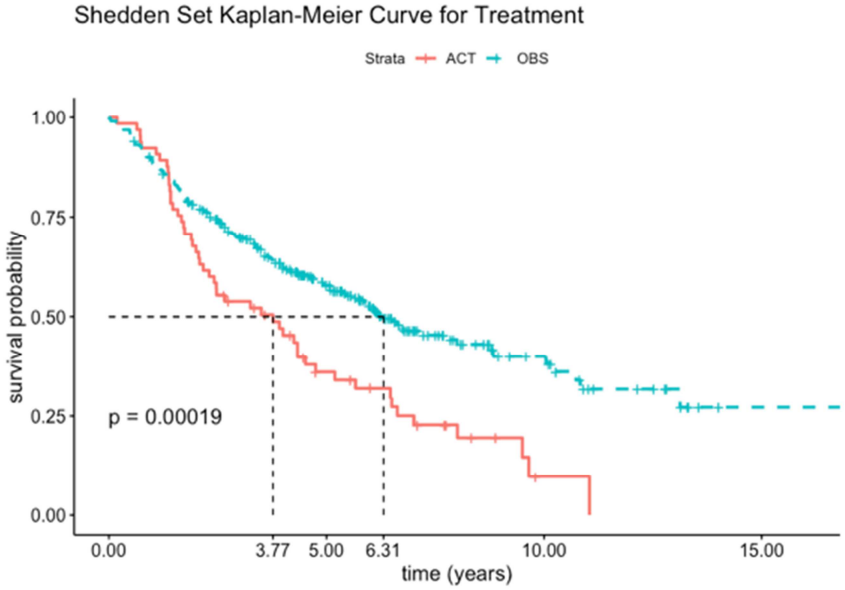

Figure 1. *Survival difference between patients who underwent OBS (dashed; top) and the patients who underwent ACT (solid; bottom) in the training data.*

Table 1. *Conversion table for TN staging to stages I, II, and III.*

Lung Cancer TN to Stages of Cancer Conversion	
	TN Stages
Stage I	01*, 02
Stage II	03, 11, 12
Stage III	04, 13, 14, 21, 22, 23, 24

*xy denotes pNxpTy.

Table 2. *Side-by-side patient information for training set and validation set.*

	Training Set ($n = 386$)	**Validation set ($n = 133$)**
Treatment Received		
Adjuvant chemotherapy (ACT)	65	71
Surgery only (OBS)	321	62
Age		
Less than 65	176	87
Older than or equal to 65	210	46

3. Methods

Survival analysis is a statistical methodology for analyzing the expected duration of time until an event of interest occurs, such as death. A main objective of survival analysis may be to estimate the survival probability that a patient can survive for a certain period of time, say five years. In

addition, survival analysis can address survival differences between treatment groups—in our example, OBS and ACT. Furthermore, important genomic biomarkers that can prolong patients' survival can be found in survival analysis by using the Cox proportional hazard model [16].

The Cox proportional hazards model is a type of regression model used in survival analysis for investigating the relationship between patients' survival and predictor variables, such as probe sets in microarray data. For a subject i, denote Z_i the value of the covariate vector $\boldsymbol{Z}$. Let T_i and C_i denote the underlying survival time and a censoring time, respectively. Let (Z_i, U_i, δ_i) be the survival data, where $U_i = \min\{T_i, C_i\}$ and $\delta_i = I_{[T_i \le C_i]}$, and where T_i and C_i are conditionally independent given Z_i. We want to model a relationship between Z and T by assuming a function $h(\cdot)$ is functionally related to Z.

Let $h(t|Z)$ denote the hazard function for a subject with covariate Z such that

$$h(t|Z) = h_0(t) * g(Z),$$

where $h_0(t)$ is a function of time t, but not Z, and $g(Z)$ is a function of Z, but not t. The Cox proportional hazard model is a special case such that

$$h(t|Z) = h_0(t) * \exp(\beta' Z) = h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i z_i\right).$$

In Cox regression, the regression coefficients are estimated by maximizing a quantity known as the partial likelihood rather than a full likelihood. In a partial likelihood, we utilize the probability of subjects (patients) who experience an event of interest rather than utilizing the whole data. Thus, patients who are censored contribute only to the risk set instead of contributing to the partial likelihood. Therefore, the likelihood takes the form

$$\mathcal{L}_p(\beta) = \prod_{i=1}^{k} L_i,$$

where k represents the number of failure times, and L_i represents a partial likelihood comparing the risk of the failed subjects with z_i to the risk given all other z_i's for subjects in the risk set at time t_i. Cox (1972) proposed the partial likelihood to estimate the parameters β

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' Z_i)}{\sum_{j \in R(t_i)} \exp(\beta' Z_j)} \right\}^{\delta_i},$$

where $R(t_i)$ is the risk set at time t_i denoting the set of individuals who are at risk for failure at time t_i [16]. The parameters β are estimated by maximizing the partial likelihood $L(\beta)$ by a Newton-Rapson technique.

In microarray data, the number of covariates is larger than the number of subjects. This type of data easily leads to overfitting and high variance problems. In this case, it is necessary to consider variable selection. We take the lasso approach for variable selection methods [17]. Later, the least squares regression models were extended into the context of Cox models by minimizing the partial likelihood with the lasso penalty [18]. The lasso estimates for the Cox model are

$$\hat{\beta}_{lasso} = \min_{\beta} L(\beta), \text{ subject to } \sum_{j=1}^{p} |\beta_j| < \lambda,$$

where λ is a specified penalty parameter and p is the number of covariates. The nature of the lasso constraint causes it to shrink irrelevant coefficients and takes out variables having coefficients that are exactly zero. As a result, it simultaneously reduces the estimation variance while providing a final interpretable model with a feasible set of variables. To minimize the bias in the estimation of parameters $\hat{\beta}$ and to make a sparse model, the parameter λ is chosen with the one having the minimum standard error in a leave-one-out cross-validation (LOOCV) due to the relatively small sample size. The LOOCV is a standard cross-validation approach that sets aside one observation as the test set and fits the model on the remaining observations in the training sample. Then, the model evaluation is carried out using the left-out test sample. This process repeats n times, where n is the sample size. This procedure is implemented using the R "glmnet" package to estimate the penalized coefficients.

Our goal is to predict the best treatment option (either ACT or OBS) for individual patients to prolong survival. Based on the lasso approach, we have reduced down to a feasible set of markers that are highly correlated with patient survival. To measure the treatment effect on the survival between OBS and ACT, a modified Cox model is trained as

$$h(t|\boldsymbol{Z}) = h_0(t) \exp(\beta_1 age + \boldsymbol{\beta}'(\boldsymbol{Z} * \boldsymbol{G})),$$

where $\boldsymbol{Z}$ includes stage of patients and the selected probe set by the lasso, and $\boldsymbol{G}$ is the treatment vector that includes either ACT or OBS depending on which one the patients in the training set have taken.

After training the model, the estimated parameters $\widehat{\boldsymbol{\beta}}$ are used to find the treatment risk scores $\widehat{\beta_1} age + \widehat{\boldsymbol{\beta}}' \boldsymbol{Z}$, that is, a risk with respect to the ACT treatment, in order to make predicted treatment recommendations for individual patients. For a future individual patient, the model examines the risk with OBS and ACT in order to stratify the individual patient into a low-risk and high-risk group. If a patient has a lower ACT risk, the model recommends ACT treatment for the patient. However, if ACT risk is higher, the model recommends OBS treatment for the patient.

To evaluate the proposed classification model, the separate test set is used to classify patients into two groups: one group of patients who are concordant with the model recommendation and the other group of patients who are discordant with the model recommendation [14]. This can be done by examining the two risk scores from the viable treatments of ACT and OBS. Then, the patient survivals from the group concordant with the model recommendation and the other discordant with the model recommendation are compared via the Kaplan-Meier survival estimator.

4. Results

The regularized Cox regression model is implemented using the training set in order to select a feasible set of probe sets. First, leave-one-out cross-validation (LOOCV) is performed for the estimation of a lasso penalty tuning parameter λ. The lasso penalty estimate with minimum standard error $\hat{\lambda}_{min}$ is used to select genomic markers for the predicted treatment recommendation. Figure 2 displays the elbow chart showing the standard error in the estimation of λ.

Table 3 displays the genomic markers that were selected by the lasso penalty using $\hat{\lambda}_{min}$. All of the selected probe sets showed a statistically significant treatment effect. These 21 genomic markers, in addition to patient age and clinical stage, were used for the estimation of the risk of taking ACT in the proposed regularized Cox regression model. Based the estimated risk, the proper treatment recommendation (either ACT or OBS) was made.

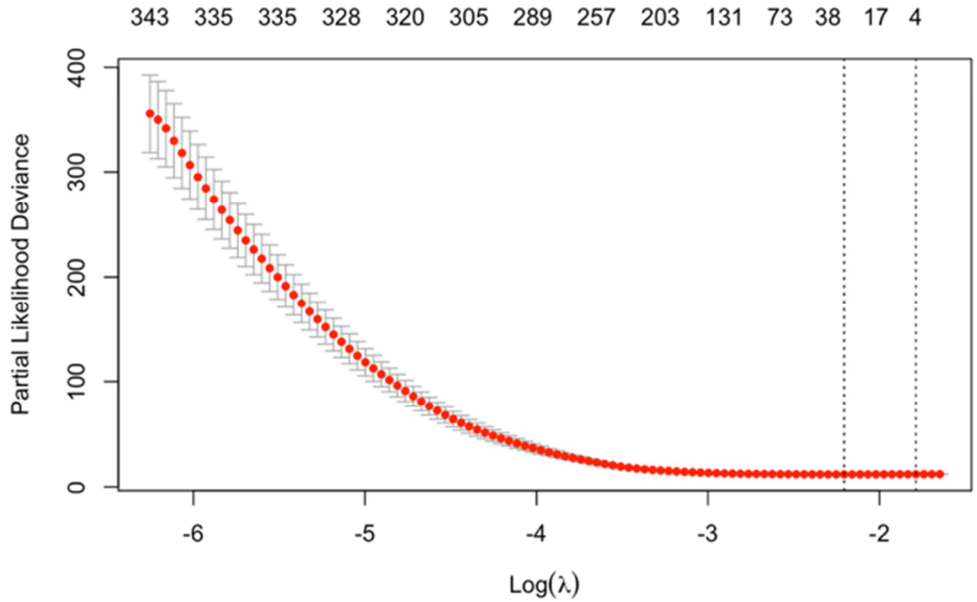

Figure 2. *The standard error is at a minimum at the left vertical line (from where the log λ value is extracted and λ is calculated), and the right vertical line shows the standard error within one standard deviation of the minimum standard error.*

Many of the selected genes in Table 3 have previously been shown to be related to NSCLC. Recently published studies suggest that CDC42 and ETV5 are associated with NSCLC oncogenesis [19, 20]. CSRP1 was found to be one of six genes consistently in the top genes for aberrant expression for NSCLC [21]. FAM164A has been previously identified as part of a 12-gene signature for lung cancer [22]. FAM117A was identified as having a possible association with lung cancer progression [23]. FOSL2 expression levels were found to be associated with reduced survival and metastases for NSCLC [24]. A recent study also found that MFHAS is important both for the progression and initiation of squamous cell lung cancer, one type of NSCLC [25]. PLEK2 was also demonstrated to be responsible for the degradation of SHIP2, which thereby allows PLEK2 to regulate vascular invasion and metastasis for NSCLC [26]. It has also been suggested that lower expression of ZNF185 is often a feature of lung tumors and may play a role in lung carcinogenesis [27]. Higher expression of FZD2 was found to correlate with longer survival and better prognosis [28]. PTPN12 expression for NSCLC patients was previously found to be a prognostic biomarker, where higher PTPN12 expression in tumors suggested longer survival [29]. In a recent study, TMPRSS11E was shown to promote the growth of lung cancer by strengthening glycolytic metabolism and the exportation of lactate [30]. CPM was found to display increased activity levels in lung cancer patients, specifically in the bronchoalveolar lavage [31].

For the predicted treatment recommendation for each individual patient, the risk score from the proposed model was used to recommend ACT or OBS (surgery only) based on which treatment option had a lower risk score for the individual. Specifically, patients who had a lower risk score for taking ACT than surgery only were recommended ACT by the proposed model and vice versa. Applying this method to the training data, 89 patients were recommended ACT, while the remaining 297 patients were recommended OBS. The patients were followed up whether they had the recommended treatment by the model for the purpose of survival analysis: one group in which the patients followed the recommended treatment and the other group in which the patients did not follow the recommended treatment. Accordingly, 268 patients who followed the recommended treatment were classified into the first group; 118 patients who did not follow the recommended treatment were classified into the second group. A log-rank test indicated that those who followed the treatment recommendation displayed a significantly higher rate of survival than those who did not follow the treatment recommendation ($p < 0.0001$; Figure 3) as we expected. Those who followed the treatment recommendation had a median survival of 7.14 years, while those who didn't follow the treatment recommendation had a median survival of 3.33 years.

The method was also employed for the JBR. 10 validation data—an unseen data set. Out of the 133 patients in the validation set, 86 patients were recommended ACT, while the other 47 patients were recommended OBS by the proposed model. When the treatment recommendations were compared to the actual treatment that each individual patient underwent, the results indicated that 68 patients followed the recommended treatment and 65 patients did not follow the recommended treatment. A significantly higher survival rate for the group of patients who followed the recommended treatment was clearly evident when comparing the two survival probabilities in Figure 4 ($p = 0.0056$). Those who followed the treatment recommendation had a median survival of at least 9.03 years, while those who did not follow the treatment recommendation had a median survival of 6.66 years. Because these results were obtained on the validation set, which the model was not trained on, the results provided supporting evidence of the model efficacy in recommending a proper treatment right after surgery for NSCLC patients. We could conclude that patients who followed the treatment recommendation provided by our proposed model could have survival benefits compared to patients who did not follow the recommendation by the model.

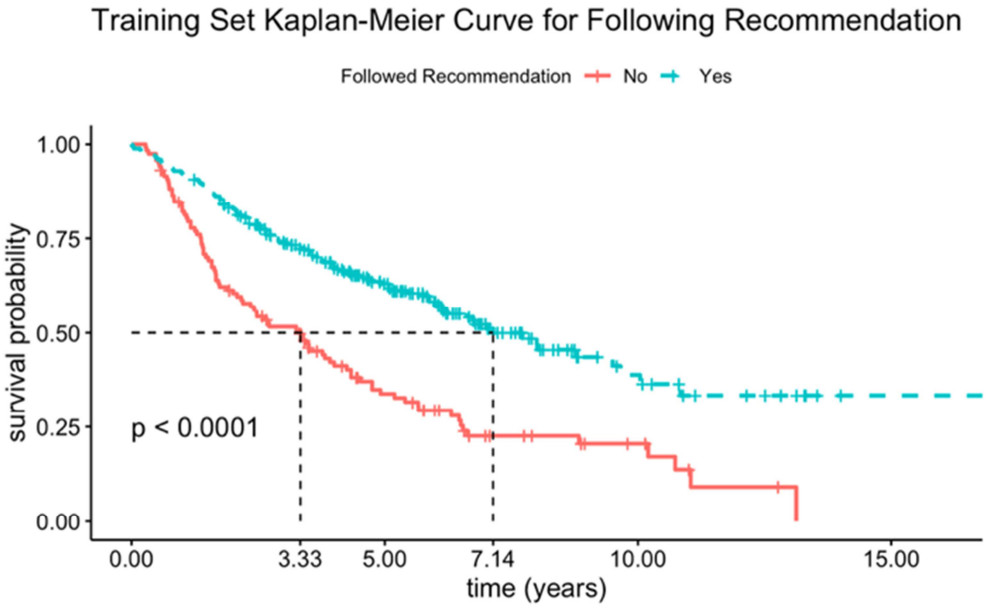

Figure 3. *Survival difference between patients who followed the treatment recommendation (dashed; top) and the patients who did not follow the treatment recommendation (solid; bottom) from the proposed regularized Cox regression model in the training data.*

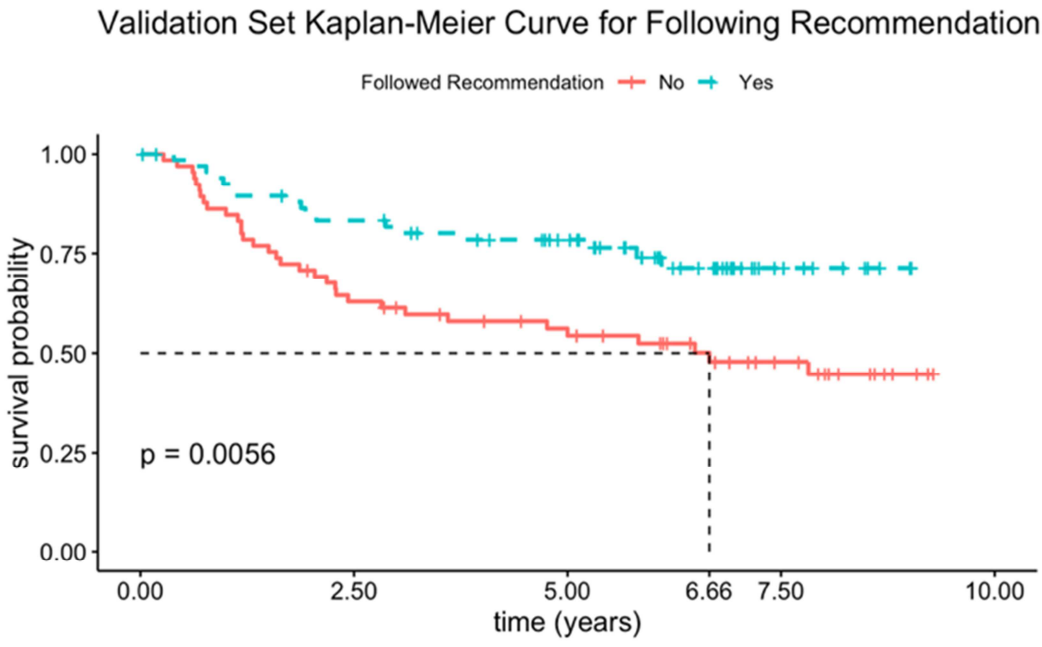

Figure 4. *Survival difference between the patients who followed the treatment recommendation (dashed; top) and the patients who did not follow the treatment recommendation (solid; bottom) from the proposed regularized Cox regression model in the JBR. 10 validation set.*

Table 3. *Genomic markers selected by the Cox regression regularized by the lasso penalty.*

Probe Set/Covariate	Gene Symbol	Gene Name
208728_s_at	CDC42	cell division cycle 42
200621_at	CSRP1	cysteine and glycine-rich protein 1
205308_at	FAM164A	family with sequence similarity 164, member A
206269_at	GCM1	glial cells missing transcription factor 1
203348_s_at	ETV5	ETS variant transcription factor 5
218885_s_at	GALNT12	polypeptide N-acetylgalactosaminyltransferase 12
218498_s_at	ERO1A	endoplasmic reticulum oxidoreductase 1 alpha
221249_s_at	FAM117A	family with sequence similarity 117 member A
209460_at	ABAT	4-aminobutyrate aminotransferase
218880_at	FOSL2	fos-related antigen 2
207629_s_at	ARHGEF2	rho guanine nucleotide exchange factor 2
213457_at	MFHAS1	malignant fibrous histiocytoma amplified sequence 1
218644_at	PLEK2	pleckstrin 2
203585_at	ZNF185	zinc finger protein 185 with LIM domain
210220_at	FZD2	frizzled class receptor 2
220183_s_at	NUDT6	nudix hydrolase 6
202006_at	PTPN12	Tyrosine-protein phosphatase non-receptor type 12
202801_at	PRKACA	protein kinase cAMP-activated catalytic subunit alpha
220431_at	TMPRSS11E	transmembrane serine protease 11E
206496_at	FMO3	flavin containing dimethylaniline monoxygenase 3
206100_at	CPM	carboxypeptidase M

5. Conclusion

The main goal of this research was to evaluate early-stage lung cancer patients on an individual basis and recommend the treatment that optimizes patient outcome. This was achieved by first identifying a set of treatment-related biomarkers and then using these biomarkers (along with age and stage) to build a Cox regression model to provide a treatment recommendation. The training set [13] consisted of 442 patients, of which 56 patients were removed due to missing data points, leaving 386 patients: 244 stage I patients, 77 stage II patients, and 65 stage III patients. The data set originally used TN staging for cancer stage, but they were converted to clinical cancer stages I, II, and III to maintain consistency with the validation set. The JBR. 10 validation set consisted of 133 patients, of which 62 underwent surgery only and 71 underwent adjuvant chemotherapy [14].

The raw training and validation sets were preprocessed before being statistically analyzed. The necessary steps of background correction, normalization, and summarization processes were conducted using Bioconductor's "affy" package in R. After the data was preprocessed and extra rows and missing data were removed, the "glmnet" package in R was used to perform LOOCV with the lasso regularization method on the training set. The lambda value with minimum standard error was extracted from the cross-validation result and used to identify 29 corresponding covariates. Of the 29 covariates, 8 were removed because they did not significantly contribute to the model, leaving 21 covariates. The 21 remaining treatment-related covariates, along with age and stage covariates, were used to build a final Cox regression model with treatment interaction terms. For each individual patient, a treatment was recommended based on the estimated treatment risk from the model. If ACT risk was lower than OBS risk, then ACT was recommended and vice versa. Patients were then classified into a group that followed the recommendation and a group that did not, and their survival difference was compared using the Kaplan-Meier survival estimates. Based on the results from the validation data, we may be able to conclude that if patients followed our predicted treatment recommendation, they could live longer. Many of the 21 selected genomic biomarkers from the proposed algorithm have been previously shown to be related to NSCLC, such as their expression levels being correlated with survival outcome or having been previously identified in a gene signature.

For the training set, the model recommended ACT to 89 patients and recommended OBS to the remaining 297 patients. Out of the total 386 patients, 268 patients followed either ACT or OBS treatment recommendation, while the remaining 118 patients did not follow the recommended treatment. The Kaplan-Meier survival estimates demonstrated that those who followed the recommendation from the model had a significantly higher survival rate (median survival of 7.14 years) than those who did not (median survival of 3.33 years) ($p < 0.0001$). These results indicate that the model performs effectively on the training data as expected.

For the validation set, the model recommended ACT to 86 patients and OBS to the remaining 47 patients. Out of the 133 patients in the set, 68 patients followed the recommended treatment and 65 patients did not. Upon comparing the two groups, the Kaplan-Meier survival estimates indicated that those who followed the treatment recommendation had a significantly higher survival rate (median survival of at least 9.03 years) than those who did not (median survival of 6.66 years) ($p = 0.0056$). The results on the validation set demonstrated the validity of the individualized treatment decisions for NSCLC patients.

We believe that our proposed model may aid physicians in making more informed treatment decisions for NSCLC patients. In particular, the model may help to improve patient

outcomes and prognosis by recommending the treatment that best suits each of the patients given their unique set of genes, age, and stage. Future research may consider immunotherapy as an additional treatment option. Immunotherapy has shown much promise in recent years, and it is believed that it will be an increasingly important and viable treatment option for cancer patients in the future. Research indicates that personalized combination therapy will be a promising cancer treatment strategy. Building a model with this additional treatment possibility would be futuristic and may be of great benefit to physicians [32].

Acknowledgements

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

References

[1] *Cancer*. (2021, March 3). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] Artal-Cortés, Á., Calera-Urquizu, L., and Hernando-Cubero, J. (2015). Adjuvant chemotherapy in non-small cell lung cancer: state-of-the-art. *Translational Lung Cancer Research*, 4 (2), 191–197.

[3] Winton, T., Livingston, R., Johnson, D., Rigas, J., Johnston, M., Butts, C., Cormier, Y., Goss, G., Inculet, R., Vallieres, E., Fry, W., Bethune, D., Ayoub, J., Ding, K., Seymour, L., Graham, B., Tsao, M. S., Gandara, D., Kesler, K., Demmy, T., and Shepherd, F. (2005). Vinorelbine plus Cisplatin vs. Observation in Resected Non-Small-Cell Lung Cancer. *The New England Journal of Medicine*, 352 (25), 2589-2597.

[4] Douillard, J. Y., Rosell, R., De Lena, M., Carpagnano, F., Ramlau, R., Gonzáles-Larriba, J. L., Grodzki, T., Pereira, J. R., Le Groumellec, A., Lorusso, V., Clary, C., Torres, A. J., Dahabreh, J., Souquet, P. J., Astudillo, J., Fournel, P., Artal-Cortes, A., Jassem, J., Koubkova, L., His, P., Marcello, R., and Hurteloup, P. (2006). Adjuvant vinorelbine plus cisplatin versus observation in patients with completely resected stage IB-IIIA non-small-cell lung cancer (Adjuvant Navelbine International Trialist Association [ANITA]): a randomised controlled trial. *The Lancet Oncology*, 7 (9), 719-727.

[5] Ibrahim, N. E. and Januzzi, J. L., Jr. (2018). Established and Emerging Roles of Biomarkers in Heart Failure. *Circulation Research*, 123 (5), 614-629.

[6] Wang, T. J. (2011). Assessing the role of circulating, genetic, and imaging biomarkers in cardiovascular risk prediction. *Circulation*, 123 (5), 551–565.

[7] Villalobos, P., and Wistuba, I. I. (2017). Lung Cancer Biomarkers. *Hematology/Oncology Clinics of North America*, 31 (1), 13–29.

[8] He, R. and Zuo, S. (2019). A Robust 8-Gene Prognostic Signature for Early-Stage Non-small Cell Lung Cancer. *Frontiers in Oncology*, 9, 693.

[9] Zuo, S., Wei, M., Zhang, H., Chen, A., Wu, J., Wei, J., and Dong, J. (2019). A robust six-gene prognostic signature for prediction of both disease-free and overall survival in non-small cell lung cancer. *Journal of Translational Medicine*, 17, 152.

[10] Boutros, P. C., Lau, S. K., Pintilie, M., Liu, N., Shepherd, F. A., Der, S. D., Tsao, M., Penn, L. Z., and Jurisica, I. (2009). Prognostic gene signatures for non-small-cell lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (8), 2824-2828.

[11] Moon, H., Zhao, Y., Pluta, D., and Ahn, H. (2018). Subgroup Analysis Based on Prognostic and Predictive Gene Signatures for Adjuvant Chemotherapy in Early-Stage Non-Small-Cell Lung Cancer Patients. *Journal of Biopharmaceutical Statistics*, 28 (4), 750-762.

[12] Moon, H., Chao, T., and Ahn, H. (2019). Identification of Risk Factors and Likelihood of Benefit from Adjuvant Chemotherapy for Early Stage Lung Cancer Patients. *Journal of Biopharmaceutical Statistics*, 30, 1-15.

[13] Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I,, Giordano, T. J., Misek, D. E., Chang, A. C., Zhu, C. Q., Strumpf, D., Hanash, S., Shepherd, F. A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Sharma, A., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., Conley, B., Seshan, V. E., Meyerson, M., Kuick, R., Dobbin, K. K., Lively, T., Jacobson, J. W., and Beer, D. G. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine,* 14, 822–827.

[14] Zhu, C-Q., Ding, K., Strumpf, D., Weir, B. A., Meyerson, M., Pennell, N., Thomas, R. K., Naoki, K., Ladd-Acosta, C., Liu, N., Pintilie, M., Der, S., Seymour, L., Jurisica, I., Shepherd, F. A., and Tsao, M. S. (2010). Prognostic and Predictive Gene Signature for Adjuvant Chemotherapy in Resected Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology*, 28 (29), 4417-4424.

[15] Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23 (20), 2700-2707.

[16] Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

[17] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58 (1), 267-288.

[18] Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, 16 (4), 385–395.

[19] Xiao, X. H., Lv, L. C., Duan, J., Wu, Y. M., He, S. J., Hu, Z. Z., and Xiong, L. X. (2018). Regulating Cdc42 and Its Signaling Pathways in Cancer: Small Molecules and MicroRNA as New Treatment Candidates. *Molecules*, 23 (4), 787.

[20] Zhang, Z., Newton, K., Kummerfeld, S. K., Webster, J., Kirkpatrick, D. S., Phu, L., Eastham-Anderson, J., Liu, J., Lee, W. P., Wu, J., Li, H., Junttila, M. R., and Dixit, V. M. (2017). Transcription factor Etv5 is essential for the maintenance of alveolar type II cells. *Proceedings of the National Academy of Sciences of the United States of America*, 114 (15), 3903–3908.

[21] Taguchi, Y. H. (2014). Integrative Analysis of Gene Expression and Promoter Methylation during Reprogramming of a Non-Small-Cell Lung Cancer Cell Line Using Principal Component Analysis-Based Unsupervised Feature Extraction. *Intelligent Computing in Bioinformatics*.

[22] Wan, Y. W., Sabbagh, E., Raese, R., Qian, Y., Luo, D., Denvir, J., Vallyathan, V., Castranova, V., and Guo, N. L. (2010). Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction. *PLOS One*, 5 (8), e12222.

[23] Wu, C. H. and Hwang, M. J. (2019). Risk stratification for lung adenocarcinoma on EGFR and TP53 mutation status, chemotherapy, and PD-L1 immunotherapy. *Cancer Medicine,* 8 (13), 5850–5861.

[24] Yin, J., Hu, W., Fu, W., Dai, L., Jiang, Z., Zhong, S., Deng, B., and Zhao, J. (2019). HGF/MET Regulated Epithelial-Mesenchymal Transitions And Metastasis By FOSL2 In Non-Small Cell Lung Cancer. *OncoTargets and Therapy,* 12, 9227–9237.

[25] Kang, J. (2015). Genomic alterations on 8p21-p23 are the most frequent genetic events in stage I squamous cell carcinoma of the lung. *Experimental and Therapeutic Medicine*, 9, 345-350.

[26] Wu, D. M., Deng, S. H., Zhou, J., Han, R., Liu, T., Zhang, T., Li, J., Chen, J. P., and Xu, Y. (2020). PLEK2 mediates metastasis and vascular invasion via the ubiquitin-dependent degradation of SHIP2 in non-small cell lung cancer. *International Journal of Cancer*, 146 (9), 2563-2575.

[27] Medina, P. P., Carretero, J., Ballestar, E., Angulo, B., Lopez-Rios, F., Esteller, M., and Sanchez-Cespedes, M. (2005). Transcriptional targets of the chromatin-remodelling factor SMARCA4/BRG1 in lung cancer cells. *Human Molecular Genetics,* 14 (7), 973-82.

[28] Ding, L. C., Huang, X. Y., Zheng, F. F., Xie, J., She, L., Feng, Y. Su, B. H., Zheng, D. L., and Lu, Y. G. (2016). FZD2 inhibits the cell growth and migration of salivary adenoid cystic carcinomas. *Oncology Reports,* 35, 1006-1012.

[29] Cao, X., Chen, Y. Z., Luo, R. Z., Zhang, L., Zhang, S. L., Zeng, J., Jiang, Y. C., Han, Y. J., and Wen, Z. S. (2015). Tyrosine-protein phosphatase non-receptor type 12 expression is a good prognostic factor in resectable non-small cell lung cancer. *Oncotarget,* 6 (13), 11704–11713.

[30] Updegraff, B. L., Zhou, X., Guo, Y., Padanad, M. S., Chen, P. H., Yang, C., Sudderth, J., Rodriguez-Tirado, C., Girard, L., Minna, J. D., Mishra, P., DeBerardinis, R. J., and O'Donnell, K. A. (2018). Transmembrane Protease TMPRSS11B Promotes Lung Cancer Growth by Enhancing Lactate Export and Glycolytic Metabolism. *Cell Reports,* 25 (8), 2223–2233.

[31] Dragović, T., Schraufnagel, D. E., Becker, R. P., Sekosan, M., Votta-Velis, E. G., and Erdös, E. G. (1995). Carboxypeptidase M activity is increased in bronchoalveolar lavage in human lung disease. *American Journal of Respiratory and Critical Care Medicine,* 152 (2), 760–764.

[32] Zhang, H. and Chen, J. (2018). Current status and future directions of cancer immunotherapy. *Journal of Cancer*, 9 (10), 1773-1781.