OPEN ACCESS

University of Dundee

University of Dundee

A two-step feature selection procedure for relevant markers of Squamous Cell Lung Carcinoma using different survival models

Bhattacharjee, Atanu; Basak, Samudranil; Kumari, Pragya

DOI:
10.1016/j.health.2023.100168

Publication date:
2023

Licence:
CC BY-NC-ND

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Discovery Research Portal

Citation for published version (APA):
Bhattacharjee, A., Basak, S., & Kumari, P. (2023). A two-step feature selection procedure for relevant markers of Squamous Cell Lung Carcinoma using different survival models. *Healthcare Analytics*, *3*, Article 100168. https://doi.org/10.1016/j.health.2023.100168

General rights
Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 11. May. 2024

Healthcare Analytics 3 (2023) 100168

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A two-step feature selection procedure for relevant markers of Squamous Cell Lung Carcinoma using different survival models

Atanu Bhattacharjee [a], Samudranil Basak [b,*], Pragya Kumari [c]

[a] Leicester Real World Evidence Unit, University of Leicester, Leicester, LE1 7RH, United Kingdom
[b] Department of Statistics, Pondicherry University, Kalapet, 605014, Pondicherry, India
[c] Department of Mathematics and Computing, Indian Institute of Technology (ISM) Dhanbad, Dhanbad, 826004, Jharkhand, India

ARTICLE INFO

Keywords:
Lung Cancer
Feature selection
High-dimensional
Lasso Cox Model
Cox Proportional Hazard Model
Accelerated Failure Time Model

ABSTRACT

There are potentially infinite gene expression markers for Lung Squamous Cell Carcinoma. This results in a high-dimensional data with a large number of features. The selection of relevant markers for analysis is thus, of utmost importance. In our study, we have aimed to select a subset of prominent and significant features from 31918 features of gene expressions. Analysis is then performed on the selected features using the Cox Proportional Hazards Model to know how each marker affects the survival estimates of a patient. We have employed a two-step selection process to select a subset of markers. The first step is done by L1 regularized Cox PH. Then the selected markers are screened a second time by running a univariate Cox PH model and checking for the p-value of each bio-marker via Wald inference ($p < 0.05$). Once the final selection is made, we estimate the Hazard Ratio and Confidence intervals using Maximum Likelihood Estimates (MLE) and the Bayesian Approach with the Cox Proportional Hazards Model (CPH) and the Accelerated Failure Time Model (AFT) as an alternative. A forest plot has also been generated to show the graphical representation of the meta-analysis done in the study. With the proposed selection procedure we have managed to find a suitable subset out of a large number of variables available. The features selected have been analyzed and their validity has been confirmed by using survival models.

1. Introduction

Lung Cancer has been regarded to be the most common category of cancer worldwide since 1958, both in terms of incidence and mortality [1]. Generally speaking, Lung cancer, can be classified as small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). Despite the accuracy of today, there has been an underestimation of statistics portrayed. The American Cancer Society estimated that in 2022, 236,740 new cases of lung cancer were detected and 130,180 deaths are directly linked to lung cancer in both men and women. Although it is debatable approximately how many people die of Lung Cancer, WHO says that numbers have been steadily increasing since the last decade.

NSCLC is classified into Squamous Cell Lung Carcinoma (LUSC), Adenocarcinoma and Large Cell Carcinoma. Squamous Cell Lung Carcinoma also known as just Squamous Cell Carcinoma begins in the main air way, such as the right or left bronchi or in the central part of the lung. Statistics say most squamous cell lung carcinoma occur due to the concerned person's history of smoking. Squamous cell carcinoma tends to occur near the central airways of the lungs. Thereby, comparisons have been drawn between Squamous cell carcinoma and Adenocarcinoma, another type of NSCLC. As such, prognostically, Adenocarcinoma was considered to be poorer as compared to Squamous cell carcinoma [2,3]. In comparison, recently, it has been shown that there was no difference in the development of recurrence between these two; however, there was a huge difference when it came to overall survival [4]. In medical research, knowing the death of an individual due to an ongoing disease or infection helps doctors prescribe medication and perform surgeries with efficiency. Survival analysis helps us answer such questions. However, each genomic sequence can potentially have thousands of permutations, and selecting the relevant ones is crucial. The main problem at hand is that although personalized therapy for Lung Adenocarcinoma have improved it has not been the same for LUSC. Recently, FAM83B was recognized as the candidate marker through a comprehensive gene expression analysis [5]. The most important thing to point out here is that extensive research is needed to identify gene expression markers that are responsible for accelerating the death of a patient diagnosed with LUSC. Over the years, biomarker identification have been easier however the process of identification still remains exhaustive. A few methods used are Tumour and Non-tumour tissue samples, RNA isolation and microarray

* Corresponding author.
E-mail address: 20375046@pondiuni.ac.in (S. Basak).

https://doi.org/10.1016/j.health.2023.100168
Received 10 December 2022; Received in revised form 4 February 2023; Accepted 2 April 2023

2772-4425/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

procedures, microarray data analysis, Quantitative real time RT-PCR analysis etc [6]. Identification is of utmost importance because this will lead to precise treatments and better quality of life for patients. In order to do so, the time taken for diagnosis needs to be minimized and with the advancement of computational prowess, Machine Learning has been the leading candidate since the early 2010's.

In cancer research, survival analysis has helped predict the probability that the patient will survive with or without intervention. By knowing the probability of the patient's survival, it is easier to predict how effective the intervention of the medical team will be and thus, improve the quality of life of the patient for a certain time interval.

This has been previously discussed by inspecting the survival rate of 130 patients with non-small cell lung cancer that was left untreated [7]. Predicting the survivability depends on the Status and the survival time of records on significant bio-markers in Non-small cell lung cancer research. However, individuals might have many permutations of genes and hence, an enormous number of bio-markers. This results in a dataset that has high dimensionality. Historically, Statistical models have always supported medical research by individualizing outcome prognostication on individual variables or by estimating the effects of risk factors that are adjusted for covariates. However, theory of statistical modeling is well defined only if the set of variables is small and fixed [8]. When the number of available features becomes largely greater than the number of sample observations ($n \gg p$) analysts face the problems of "High-dimensionality".

While dealing with data in high-dimensional space, feature selection is of utmost importance. Bellman [9] coined the problems of organizing and analyzing high dimensional data sets as the "curse of dimensionality". High-dimensional data came into existence with the rise of modern technology. At present, the measurement of many variables simultaneously is possible, which has led scientists and researchers to build data sets that have far more number of covariates than the sample size. While such data sets can provide a flood of information, researchers have faced a plethora of challenges while dealing with them [10]. Pires and Branco [11] explains how whilst high dimensionality is beneficial sometimes it is regarded as one of the latest challenges faced by data analysts. Another challenge is that when the number of observations is less as compared to the number of co-variates(or features), the model runs the risk of being overfitted. Also, with the increase in the number of features, the observations become harder to cluster, and overall knowledge discovery becomes difficult.

It is thus, in the interest of analysts to extract a subset of relevant variables (or features) that are deemed to be the most relevant for the analysis. Thus, over the years variable selection methods have been proposed and applied in various fields of research including computational biology and health sciences, especially for genomic data. Variable selection methods are numerous and they all work in different algorithm schemes covering from fast correlation based filtering methods [12] to classical embedded methods [13] to more modern and complex methods such as LASSO [14], Lars [15], Sure Independence Screening [16]. Some of these methods can deal with high-dimensionality quite well while others cannot. Wasserman and Roeder [17] explores the possibilities of the above-mentioned variable selection methods in high-dimensional models.

Whilst there are an extensive number of machine learning algorithms that are quite useful for feature selection, no method is considered to be the best. It all comes down to the data set we have in hand. The Partial likelihood method was prevalent for a long time for estimating parameters for the Cox model however, lately, neural network algorithms such as the Coxnnet [18] have gained popularity. In a different perspective, selection of features can also be considered to be imbalanced classification problems and gradient boosting decision tree algorithms such as the LightGBM [19] prove to be a great solution. LightGBM used as a feature selection algorithm has been previously used in studies related to predicting Drug Target interactions [20] and Phage Virion Protein classification [21]. Another efficient method, while still in their infancy has proved to be quite useful in identifying pneumonia-related compounds is the CapsNet [22,23]. CapsNet is interesting because it strongly highlights the generalization capabilities of capsules over traditional neural networks and may prove to work better in identifying relevant gene expressions as well. Needless to say neural network methods work best for classification problems but they are computationally much more exhausting.

Another interesting application that is a strong candidate for the future would be to evaluate different algorithms based on an ensembling method. Bao et al. [24] have demonstrated this with their study on identification of Lysine 2-hydroxyisobutyrylation by generating interaction ranking lists and then evaluating their performances by 3 ensembling methods. Algorithms like the VIKOR method [25] has been used as a feature evaluation method as different decision-making criteria [26], integration of multiple ranking information with an SVM ensemble model [27] and a tree-based stacked ensemble technique (SET) [28] have been previously studied and has proven to work in numerous fields. However, ensembling methods also reduces the interpretability of the model for analysis and thus, a comprehensive study is needed to test their accuracy in survival and censored data sets.

The Lasso Cox model [29] has worked wonders previously with survival and censored data sets and moreover, is a much easier alternative than advanced neural network methods. To demonstrate this, Qian et al. [30] studied the prognosis of breast cancer and incorporated the Univariate Cox PH and Lasso cox regression model to identify the 17-gene signatures. A similar study was also done by Zhang et al. [31] where the Lasso Cox regression model was utilized for constructing a prognosis prediction model for Lung adenocarcinoma.

In this study, we have thus, attempted to utilize a two-step selection procedure which includes the Lasso Cox Model [29] to select a subset of features from high-dimensional gene expression data of patients diagnosed with Squamous Cell Lung Carcinoma. The main objective of this study is to find relevant biomarkers and then validate them using 2 survival models.

Once the biomarkers are selected, we estimate the respective Hazard Ratios and their confidence intervals using both Maximum Likelihood Estimates (MLE) and Bayesian approaches using the Cox Proportional Hazards (CPH) Model and the Accelerated Time Failure (AFT) Model. The Cox proportional hazards model, developed by Sir David Cox in 1972, is a semi-parametric model used to predict overall survival on multiple predictors [32]. It has been extended in the past for analyzing known clinical prognostic variables. Herndon et al. [33] made use of it in their study of whether the quality of life is predictive of survival of patients with advanced non-small cell lung carcinoma. Estimation of the parameters for the CPH model is done via both MLE and the Bayesian approach. [34] explains how the Bayesian approach estimates parameters in the Cox model by maximizing partial likelihood functions on the basis of previously known information in their study of the HACE1 gene in the onset of Alzheimer's disease.

Alternatively, the Accelerated Failure Time model is also applied to find the regression coefficients. Unlike the CPH model, the AFT model is a parametric model and it assumes that the effect of a co-variate either will accelerate or decelerate the event time by a constant [35,36]. The AFT model is used to compare and validate our findings.

2. Models and methods

2.1. Data motivation

For this study, we have used a high-dimensional gene expression data set of 242 unique patients suffering from Lung Squamous Cell Carcinoma (LUSC) obtained from The Cancer Genome Atlas Program (TCGA), available online at https://portal.gdc.cancer.gov. The data set used has more than 31,000 unique variables providing information focused on patient history. Variables of interest are Event and Overall

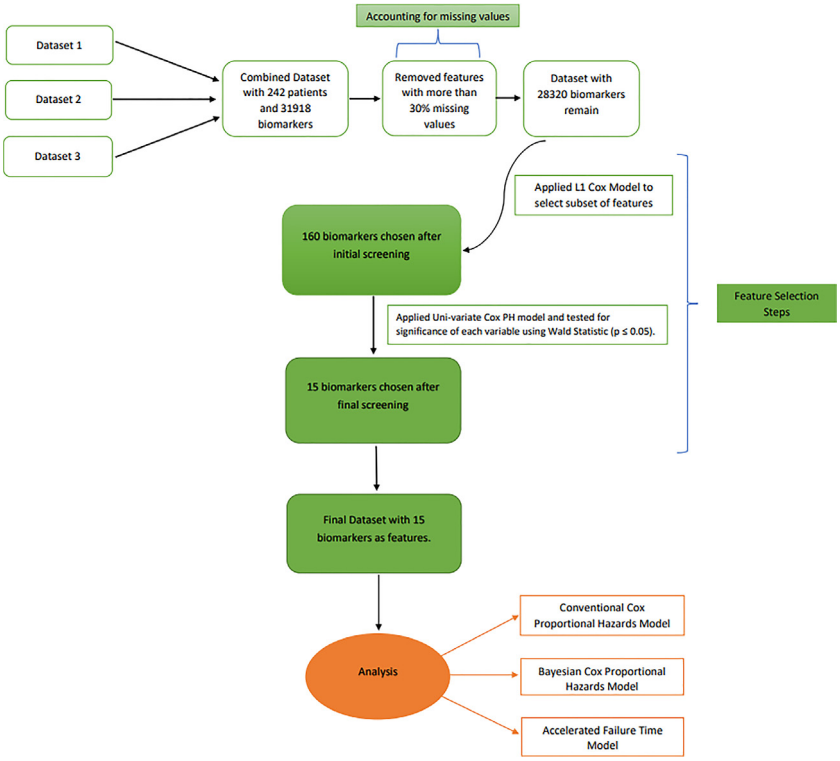

Fig. 1. Flowchart for selection of variables and analysis.

Survival (OS) along with 31,918 markers or gene expressions for each patient. The OS (in days) has been calculated for each patient which signifies the number of days the patient has lived and the Event variables shows whether that patient is alive or dead (dichotomous in nature). For our study, we consider the Event to be 1 for all patients, signifying that none of the patients have survived. Pre-processing of the data is first done by removing columns that have more than 30% missing values. The total number of variables are thus found to be 31,920 which includes our variables of interest, OS and Event along with 31,918 markers or gene expressions. It is important to note, that in our considered data set, the OS time has been counted in days. Approval and consent of patients have been taken in the creation of the data set and the data set is on an open access space and is available for public research purposes.

In this paper, we have utilized a two-step filtration method for feature selection. The method is based on the Lasso Cox Model (L1 regularized Cox PH) followed by testing the significance using a univariate Cox model screening step. Variables with p-values less than 0.05 are selected for our final data set which are then analyzed. The method of Lasso has been around for a long time since its inception in the late 90's and is a reliable method for feature selection which is later discussed in the paper. A detailed flowchart of the pre-processing, selection and analysis pertaining to the study is given in Fig. 1.

2.2. Model formulation

2.2.1. Lasso Cox Model

The lasso or Least absolute shrinkage and selection operator is extremely useful in statistics and machine learning for the purpose of both variable selection and regularization. As discussed, the presence of a large number of variables in a model can makes for the model to be interpreted. It is, thus, quite clear why variable selection is beneficial in the model building stage. Tibshirani [14] addressed the problems that analysts face while dealing with the interpretability of Ordinary Least Squares (OLS) estimates. The first problem faced is that of accuracy in prediction. Least Squares estimates tend to have low bias and large variance. It has been observed that prediction accuracy improves when some coefficients are set to 0. The second problem is that of interpretation. In the presence of high number of co-variates or features, the model becomes extremely difficult to interpret.

We consider a data set consisting of n subjects. Now, let us assume that we have X = x_1, x_2, …, x_m set of possible features in our data set. Out of these m features, we need to select a set of, say, k features, $k < m$, that are deemed to be relevant. We assume Y is the variable to be predicted. Thus, the equation for the prediction can be predicted by:

$$Y = f(\beta_0, \beta, X, \epsilon) \tag{1}$$

where, β_0 is the intercept, $\beta = \beta_1, \beta_2, \beta_3, \ldots, \beta_m$ are the possible coefficients of X and ϵ is the error term. The lasso regularization methods adds a penalization factor to the maximum likelihood estimation function. The lasso function consists of a constant, say, λ_l. It is then multiplied with the absolute values of the parameter estimates. The corresponding value is thus denoted by

$$\lambda_l \Sigma_i |(\beta_i)| \tag{2}$$

[37]. Now, as we minimize the negative log-likelihood function, for i = 1,2,..,n, the LASSO function L_l is defined as,

$$L_l = L + \lambda_l \Sigma_i |(\beta_i)|. \tag{3}$$

Where L is the corresponding likelihood function. When selecting specific bio-markers for a disease, it is important that we focus on the significant ones and ignore the rest. It is already known that L2 regularization fails to do that. This is because it reduces the impact of each factor on the model but it does not eliminate the factor altogether. Researchers thus have preferred LASSO whenever they need to ignore certain factors or variables in a data set, especially one that is High-Dimensional. L2 regularization can be used when the analyst has to include all the variables present in the data set.

The data setup is considered to be in the usual survival form. The data is thus expressed in the form of: $(y_1, x^1, \delta_1), \ldots, (y_N, x^N, \delta_N)$, where

y_i is the survival time completed, if $\delta_i = 1$ and is right censored if $\delta_i = 0$, with x^i being the vector of predictors [29].

The cox model is expressed as:

$$\lambda(t|x) = \lambda_0(t)e^{\Sigma_j x_j \beta_j} \tag{4}$$

where, t is the survival time and $\lambda(t|x)$ is the hazard function explained by a group of predictors $x = (x_1, x_2, \ldots, x_p)$. In order to estimate the coefficients, β, we usually maximize the partial likelihood without the specification of $\lambda_0(t)$. The partial likelihood function is defined by:

$$L(\beta) = \prod_{i=a}^{b} f(i)\frac{e^{s_i\beta}}{[\Sigma_{m\in(t_i)}e^{z_m\beta}]^d i} \tag{5}$$

where, $s_i = \Sigma z_i q$ is the sum of co-variates of the item observed at failure time t_i. The value of β can be found by maximizing the Eq. (3). A detailed explanation can be found in [29]. In the proportional hazards model with L1 penalization, the coefficient vector β, is estimated via the criterion

$$\hat{\beta} = argminl(\beta) \tag{6}$$

subjected to $\Sigma|\beta_j| \leq$ s. Here, s is a user-specified parameter. We consider the maximizers of the partial likelihood to be $\hat{\beta}_j^0$. It thus follows that if s $\geq \Sigma|\hat{\beta}_j^0|$, the solutions to [6] are the partial likelihood estimates. Else if s $< \Sigma|\hat{\beta}_j^0|$ then the estimates are shrunken to zero. In other words, unlike the likelihood estimation, we use a parameter s, in order to determine the coefficients, the following algorithm is followed.

The one-term Taylor series expansion is given as:

$$(z-\eta)^T A(z-\eta) \tag{7}$$

The procedure we used from [29] for obtaining estimates of the parameters of the PH model via Lasso is:

1: Fix value of s and $\hat{\beta} \leftarrow 0$.
2: Compute η, u, A and z based on the current value of $\hat{\beta}$.
3: Minimize $(z - X\beta)^T$ A $(z - X\beta)$ subject to $\sum|\beta_i| \leq$ s.
4: Repeat steps 2 and 3 until the value of $\hat{\beta}$ does not change.

2.2.2. Cox Proportional Hazards Model

The Cox Proportionals Hazards Model (CPH) is well known in the field of Survival Analysis and Biostatistics. It is a semi-parametric survival model that relates the time that passes before which an event occurs. In a CPH model the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. It was developed by Sir David Cox in 1972 [38].

In this study, the CPH model is used as a secondary screening step in our variable selection process and also to analyze the significance of the chosen biomarkers.

The cox model is given as:

$$\lambda(t|x) = \lambda_0(t)e^{\Sigma_j x_j \beta_j} \tag{8}$$

where, t is the survival time and $\lambda(t|x)$ is the hazard function explained by a group of predictors $x = (x_1, x_2, \ldots, x_p)$. In order to estimate the coefficients, β, we usually maximize the partial likelihood without the specification of $\lambda_0(t)$. The performance of the selected markers is analyzed with the Hazard Ratio (HR). The Hazard ratio for two co-variates is defined by:

$$HR = \frac{h_{y1}(t)}{h_{y2}(t)} = \frac{e^{y_1\beta}}{e^{y_2\beta}} \tag{9}$$

where, y_1(t) and y_2t are the two co-variates. The Hazard Ratio, thus, quantifies the measure of difference between the two groups of co-variates. The likelihood of occurrence of the event or risk increases if $HR > 1$ by $(HR - 1) \times 100\%$ and it decreases if $HR < 1$ by (1-HR)× 100%. If $HR = 1$, there is a lack of association [37].

In addition to utilizing the Proportional Hazards Model via Maximum Likelihood Estimation, a Bayesian approach is also executed. This estimation procedure is based on prior information about the data set. The advantage of using Bayesian Survival Analysis(BSA) is that the Bias tends to be very less, and the standard error is also far smaller than the Cox Regression Analysis regardless of the sample size. Unlike the Maximum Likelihood approach, where inferences are drawn based on the likelihood function of the data, the Bayesian model considers the likelihood to be a function of a set of parameters β given the co-variates x_i. Let us consider the likelihood of the given observations x given a set of parameters β as $p(x|\beta)$. We also consider the prior information density is given by $\pi(\beta)$. Therefore, the simple relationship between the densities would be given by:

$$p(\beta|x) \propto p(x|\beta) \times \pi(\beta) \tag{10}$$

The need to generate samples to know the updated information of the parameters β is crucial. Generating samples can be done by Markov Chain Monte Carlo (MCMC) simulation where the sample is generated from an underlying target distribution most commonly the Normal Distribution or maybe a mixture of several distributions. It is important to note that prior information is vital in BSA. The regression coefficients are considered to be the parameters in the Proportional Hazards Model [37]. The estimates using the Bayesian approach are found by using the "SurvMC" function in the "SurviMChd" package [39] in R with 10000 iterations.

2.2.3. Accelerated Failure Time Model

The Accelerated Failure Time (AFT) model is sometimes used as a substitute to the Proportional Hazards Model. The AFT is a parametric model as opposed to the semi-parametric Cox PH model, and it is used to define the relationship among the response and the survival time. The AFT model assumes that the effect of the co-variates act proportionally with respect to the survival time, which a stark contrast from the CPH model. We consider, T_i to be the failure time for the ith patient, i = 1, …,n. We consider x_i to be a p×1 of co-variates for T_i. Now, $log_{10}(T_i)$ is linearly related to x_i and there exists a constant theta, such that,

$$logT_i = \theta x_i + \epsilon_i \tag{11}$$

Therefore, the equation for the AFT model can be expressed as:

$$\lambda(t|x) = \theta\lambda_0(\theta t) \tag{12}$$

The assumption of the AFT model can also be expressed as:

$$s(t|x) = s_0(e^{(\beta' x)t}) \tag{13}$$

where, s(t|x) denotes the survival function and $s_0(e^{(\beta' x)t})$ denotes the baseline survival function at time t. The factor by which the survivability of a patient increases or decreases is $e^{(\beta' x)}$ and is known as the acceleration factor. Thus, unlike the CPH model, the AFT model gives us the effect of covariates that proportionally acts with respect to the survival time [35,36]. We can interpret the coefficients estimated by considering the unit increase in covariates will increase the mean (or median) survival time by e^{β}. Therefore, if the coefficient is positive, then the $e^{\beta} > 1$, will decelerate the event time and increase the mean (or median) survival time. On the contrary, if the coefficient is negative, then the $e^{\beta} < 1$ accelerates the event time and decreases the mean (or median) survival time. The estimates of the AFT model is obtained using the "rglaft" function in the "afthd" package [40] in R.

3. Results

The data set we have used for this study consists of 31918 gene expressions along with Overall Survival (OS) and Event status for 242 unique patients. After the variable selection process, the total number of variables is 17, which includes OS, Event, and 15 gene expression markers.

We have used L1 regularized Cox PH for the initial screening of the markers. The L1 regularized Cox PH procedure is given in the previous section. After the initial screening, we got 160 variables that

Table 1
Selected markers with respective gene names.

Probe ID	Gene name
ENSG00000099860_7	GADD45B
ENSG00000118515_10	SGK1
ENSG00000125503_11	PPP1R12C
ENSG00000165424_6	ZCCHC24
ENSG00000182325_9	FBXL6
ENSG00000185168_5	LINC00482
ENSG00000196295_10	AC005154.6
ENSG00000204967_9	PCDHA4
ENSG00000221571_3	RNU6ATAC35P
ENSG00000250995_1	RP13
ENSG00000259083_1	RP11
ENSG00000259954_1	IL21R-AS1
ENSG00000269836_1	CTD-3032J10.4
ENSG00000270890_1	RP3-468K18.6
ENSG00000276570_1	CTD

Table 2
Estimates of Hazard Ratio using the Cox proportional hazard model.

Variable	HR	Confidence Interval (95%)	P-Value
GADD45B	1.12	(1.02, 1.22)	0.01
SGK1	1.12	(1.01, 1.24)	0.04
PPP1R12C	1.21	(1.05, 1.39)	0.00
ZCCHC24	1.10	(1.01, 1.20)	0.03
FBXL6	1.21	(1.05, 1.38)	0.00
LINC00482	1.09	(1.02, 1.17)	0.00
AC005154.6	1.18	(1.02, 1.37)	0.02
PCDHA4	1.07	(1.02, 1.14)	0.01
RNU6ATAC35P	0.88	(0.79, 0.99)	0.03
RP13	1.24	(1.11, 1.39)	0.00
RP11	1.13	(1.01, 1.28)	0.03
IL21R-AS1	1.13	(1.02, 1.26)	0.02
CTD-3032J10.4	1.16	(1.03, 1.31)	0.01
RP3-468K18.6	1.13	(1.01, 1.26)	0.03
CTD	1.15	(1.02, 1.31)	0.01

Table 3
Posterior estimates of Hazard Ratio(HR) in Cox proportional hazard model.

Variable	Mean	SD	HPD interval
GADD45B	1.12	0.05	(1.02, 1.21)
SGK1	1.12	0.06	(1.00, 1.24)
PPP1R12C	1.21	0.08	(1.05, 1.38)
ZCCHC24	1.10	0.05	(1.00, 1.19)
FBXL6	1.21	0.08	(1.04, 1.37)
LINC00482	1.09	0.03	(1.02, 1.17)
AC005154.6	1.19	0.08	(1.02, 1.37)
PCDHA4	1.07	0.03	(1.01, 1.13)
RNU6ATAC35P	0.88	0.04	(0.79, 0.98)
RP13	1.24	0.07	(1.10, 1.38)
RP11	1.14	0.07	(1.00, 1.28)
IL21R-AS1	1.13	0.06	(1.01, 1.25)
CTD-3032J10.4	1.16	0.06	(1.03, 1.30)
RP3-468K18.6	1.13	0.06	(1.01, 1.20)
CTD	1.16	0.07	(1.02, 1.30)

Table 4
Estimates based on the Accelerated Failure Time Model.

Variable	Estimate	P-value
RP13	−0.21	0.00
FBXL6	−0.15	0.00
LINC00482	−0.10	0.00
PPP1R12C	−0.14	0.00
CTD-3032J10.4	−0.16	0.00
CTD	−0.15	0.01
PCDHA4	−0.07	0.01
RP11	−0.14	0.01
RP3-468K18.6	−0.13	0.01
GADD45B	−0.11	0.02
AC005154.6	−0.14	0.02
RNU6ATAC35P	0.13	0.02
SGK1	−0.10	0.02
IL21R-AS1	−0.12	0.02
ZCCHC24	−0.08	0.06

were selected via the Lasso method. We then use a univariate Cox PH filtration based on the Wald test to determine the statistical significance of the markers. In the end, 15 prominent and significant markers were selected.

The selected markers with their respective Gene names are given in Table 1.

After selection of prominent markers, parameters have been estimated using both the CPH and AFT model. The CPH model has a robust nature that allows us to find estimates of a parameter using the correct parametric model.

For our analysis, we make use of both the Cox Proportional Hazards model and the Accelerated Failure Time model. The equations for both the models is given by Eqs. (5) and (8), respectively. For the PH model, we first estimate the Hazard Ratios using the conventional method, i.e., by Maximum likelihood Estimates and then by Bayesian Survival Analysis. The estimates of the Hazard Ratio calculated by the conventional method, along with their confidence intervals and P-values, are given in Table 2.

A forest plot is obtained for easily visualized interpretation of estimates obtained for Maximum Likelihood Estimates of the CPH model and shown in Fig. 2. Hazard ratios and their corresponding P-values (from Table 1) are used to get the forest plot. We have constructed a forest plot based on the results obtained in Table 2. The forest plot is a visualization tool used for meta analysis. It is also often used to summarize the effects of many variables in a single image format. The vertical axis represents the line of null effect ($HR = 1$, in our case). The horizontal axis denotes the scale for the statistics being displayed. The plot will thus show us which selected markers have a positive (or negative) impact on the event (death).

Table 2 shows that all the co-variates are more significant than 1, except for RNU6ATAC35P, which estimates 0.89. The co-variates having a Hazard Ratio (HR) greater than 1 are accountable for increasing the risk of death due to Squamous Cell Lung Carcinoma. Therefore, the higher the Hazard Ratio, the higher the chances are that a patient suffering from LUSC will die. Hence, all the bio-markers whose HR is more significant than one are given in Tables 1 and 2. For bio-markers with a value of HR > 1, those markers have a higher risk of the Event occurring than those whose value is < 1. For example, the marker GADD45B has an HR of 1.12. This means that it has a 12% more chance of causing the death of a patient related to LUSC. Thus, the co-variates that display a higher-valued hazard ratio are set to increase the risk of death. It is also noted that co-variates with hazard ratio =1 imply a lack of association.

Fig. 2 shows that PCDHA4 and LINC00482 are the co-variates with HR closest to 1. Therefore, they are less significant than the other biomarkers. Their 95% confidence intervals are also containing 1. On the contrary, when the compliment of the Hazard Ratio is less than 1, the survival chances of a patient increase. For example, for RNU6ATAC35P, the HR is 0.89. Therefore the chances of reaching death are decreased by 11% approximately.

We have used the Bayesian Approach of the CPH model to calculate the Hazard Ratios' estimates and their HPD (Higher Posterior Density) intervals, which are given in Table 3.

We also find the estimates of the coefficients using the Accelerated Failure Time model as a substitute to the Proportional Hazards Model based on the selected markers.

In Table 4, we see that most of the estimates of the regression coefficients are negative, except RNU6ATAC35P. This would imply that most of the bio-markers selected to increase the risk of the Event. Here, we consider the Event to be death. Therefore, the negative values of the estimates tell us that the event time will accelerate. For example, the forecast of the bio-marker RP13 is −0.214443609, which implies that one unit increase in the value of this bio-marker changes the mean

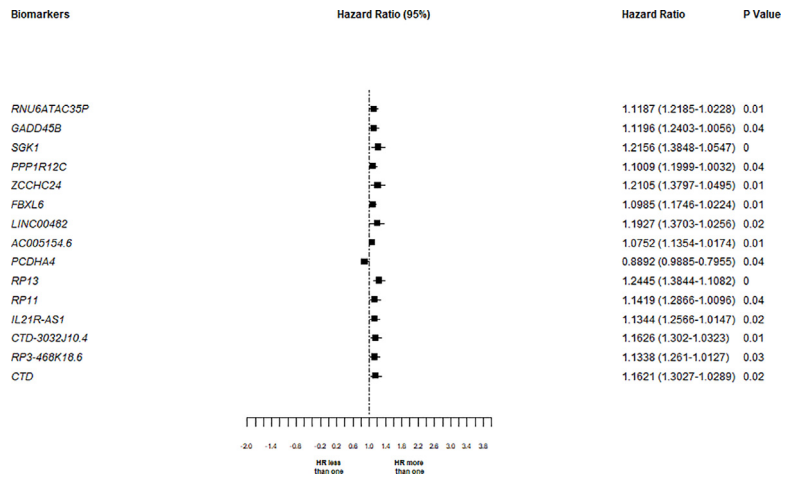

Fig. 2. Forest Plot of estimated Hazard ratios of selected markers.

survival time by a factor of 0.806986. This bio-marker accelerates the event time or reduces the mean survival time. Similarly, the bio-marker RNU6ATAC35P has an estimate of 0.13094308, which means one unit increase in the value will change the mean(or median) survival time by a factor of 1.1399028962. Therefore, we see that changes in most of these bio-markers bring forth a decrease in mean(or median) survival time except for RNU6ATAC35P.

4. Discussion

The problems of ultra-high dimensionality is infamous in the field of statistics and machine learning and has been troubling researchers for decades especially in the field of bio statistics. Diagnosis and detection of the disease at an early stage is crucial for the improvement of the quality of life of a patient. The detection of certain gene expression levels can definitely be used for diagnosis, prognosis and overall treatment [41]. In this study, we have incorporated a two-step feature selection process to filter relevant gene markers for Squamous Cell Lung Carcinoma. We incorporate the L1 regularized Cox Model for our initial feature selection and univariate Cox Model with respect to Wald inference for our second screening. Squamous Cell Lung Carcinoma data (provided by TCGA) contains a large number of gene expressions for each patient and the proposed selection method has provided adequate results that has aided in variable selection. The L1 Cox Model has been a viable option for feature selection and performs better than L2 and L3 regularization. It has already been previously demonstrated that the L1 regularization is a better variable selection method than the L2 method [42]. Vishwakarma et al. [37] has also demonstrated and stated that L1 regularization is in fact, a better option when features have to be eliminated altogether because the L2 method only shrinks the coefficient of the variables.

The initial feature selection step, thus, comprises of the L1 regularized Cox Model that does a good job in reducing the number of features from 31918 to 160. However, using all 160 features in our model is not feasible because it is still too many at once to be considered. Therefore, as a second selection step we use the univariate Cox PH model and check for the p-value (≤ 0.05) for each gene expression marker. This allows us to identify which of the markers are statistically significant. The second step lets us narrow down the features from 160 to 15. It is worth noting that the regularization methods in general does not define the biological relationships or relevance of the markers selected [12]. Thus, we have validated the significance of the biomarkers selected by performing Multivariate analysis of the CPH regression using both the conventional Cox Proportional Hazards with Maximum Likelihood Estimation and the Bayesian method. The Accelerated Failure Time model is also used as an alternative.

Ghosh and Chinnaiyan [43] have previously used Lasso for the classification and selection of genomic markers in a combination of simulated and prostate cancer gene expression data. Vasquez et al. [44] have compared the Lasso to 5 Lasso-type methods on Tucson Epidemiological Study of Airway Obstructive Disease (TESAOD) data with a group of 86 serum bio-markers and have concluded that based on the simulation study no method had any overall superiority in performance. The Lasso rightly identified more true signals and did not include noise variables more than the Weighted Fusion method. Kim and Bredel [45] have also demonstrated the Cox regression method with Lasso Optimization and found that using whole-genome gene expression data demonstrated a higher survival prediction power than other methods used such as 1-NN method but was outperformed by the same method when using gene expression profiles of cancer pathway genes alone. Despite the practicality and easiness of the L1 method, there may be some methods that are computationally more powerful like the Iterative Sure Independence screening method (ISIS) [16]. A combined Iterative Sure Independence Screening and the Cox Proportional Hazard Model study has been done to extract and analyze biomarkers for Lung Adenocarcinoma and has been proved to be effective with ultra-high dimensional data sets [46].

Overall, the Lasso Cox Model works quite well when it comes to a survival characteristic data combined with gene expression markers. The univariate Cox Proportional Hazards model acts as a secondary screening to finalize which gene markers are statistically significant in the study. However, there are a few counter arguments. It has been previously shown how L1 regularized Cox model have failed to work for highly correlated data where the method fails to select any variables at all [47]. An alternative is suggested using the Laplacian regularized Cox PH Model. Thus, this needs to be further investigated to overcome the issue of multi-collinearity.

5. Conclusion

In conclusion, the L1 regularized Cox model, combined with the univariate Cox Model, provides a beneficial method for variable selection in high-dimensional gene expression data. We have shown how the two-step feature selection method implemented can be an efficient and effective method in determining significant biomarkers from a high-dimensional data set where the number of independent variables are much larger than the sample size. Using the specified selection procedure method, we have identified 15 biomarkers that have significant impact on overall survival time in patients diagnosed with LUSC. Based on the results provided in Tables 2, 3 and 4, we observe that most of the biomarkers impact the overall survival time of the patients negatively, that is, they accelerate the time to death. It is important to note, that since the data set is censored we have confirmation that all patients

considered in the study had an Event of 1 and that implies that the biomarkers selected by the selection procedure has accurately selected the biomarkers. The statistical significance of the selected biomarkers with their p-values are displayed in the forest plot to give readers a better understanding.

It is known that there exists other machine learning algorithms that can select a subset of features from a high-dimensional data set but we have found a fast selection procedure that selects features based on the form of the constraint. However, further study needs to be done to counter multi-collinearity issues that may arise with the Lasso. It is hopeful that the selected biomarkers may play a vital role in determining significant prognostic factors of Squamous Cell Lung Carcinoma. Observing and extracting relevant biomarkers that play a significant part in a serious disease provides us a window to detect and treat the disease in early stages and improve the overall quality of life of patients.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C.S.D. Cruz, L.T. Tanoue, R.A. Matthay, Lung cancer: epidemiology, etiology, and prevention, Clin. Chest Med. 32 (4) (2011) 605–644.

[2] Y. Ichinose, T. Yano, H. Asoh, H. Yokoyama, I. Yoshino, Y. Katsuda, Prognostic factors obtained by a pathologic examination in completely resected non-small-cell lung cancer: an analysis in each pathologic stage, J. Thorac. Cardiovasc. Surg. 110 (3) (1995) 601–605.

[3] K. Suzuki, K. Nagai, J. Yoshida, M. Nishimura, K. Takahashi, T. Yokose, Y. Nishiwaki, Conventional clinicopathologic prognostic factors in surgically resected nonsmall cell lung carcinoma: a comparison of prognostic factors for each pathologic TNM stage based on multivariate analyses, Cancer 86 (10) (1999) 1976–1984.

[4] A. Kawase, J. Yoshida, G. Ishii, M. Nakao, K. Aokage, T. Hishida, M. Nishimura, K. Nagai, Differences between squamous cell carcinoma and adenocarcinoma of the lung: Are adenocarcinoma and squamous cell carcinoma prognostically equal? Jpn. J. Clin. Oncol. 42 (3) (2011) 189–195.

[5] N. Okabe, J. Ezaki, T. Yamaura, S. Muto, J. Osugi, H. Tamura, J.-I. Imai, E. Ito, Y. Yanagisawa, R. Honma, et al., FAM83b is a novel biomarker for diagnosis and prognosis of lung squamous cell carcinoma, Int. J. Oncol. 46 (3) (2015) 999–1006.

[6] A. Sanchez-Palencia, M. Gomez-Morales, J.A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell, M.E. Fárez-Vidal, Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer, Int. J. Cancer 129 (2) (2011) 355–364.

[7] E. Vrdoljak, K. Miše, D. Sapunar, A. Rozga, M. Marušić, Survival analysis of untreated patients with non-small-cell lung cancer, Chest 106 (6) (1994) 1797–1800.

[8] G. Heinze, C. Wallisch, D. Dunkler, Variable selection–a review and recommendations for the practicing statistician, Biom. J. 60 (3) (2018) 431–449.

[9] R. Bellman, Dynamic programming, Science 153 (3731) (1966) 34–37.

[10] J. Fan, R. Li, Statistical challenges with high dimensionality: Feature selection in knowledge discovery, 2006, arXiv preprint Math/0602133.

[11] A. Pires, J. Branco, High dimensionality: The latest challenge to data analysis, 2019, arXiv preprint arXiv:1902.04679.

[12] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 856–863.

[13] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1) (1997) 273–324, Relevance.

[14] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[15] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499.

[16] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5) (2008) 849–911.

[17] L. Wasserman, K. Roeder, High dimensional variable selection, Ann. Statist. 37 (5A) (2009) 2178.

[18] T. Ching, X. Zhu, L.X. Garmire, Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data, PLoS Comput. Biol. 14 (4) (2018) e1006076.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[20] Y. Zhang, Z. Jiang, C. Chen, Q. Wei, H. Gu, B. Yu, DeepStack-DTIs: predicting drug–target interactions using lightgbm feature selection and deep-stacked ensemble classifier, Interdiscip. Sci.: Comput. Life Sci. (2022) 1–20.

[21] W. Bao, Q. Cui, B. Chen, B. Yang, Phage_UniR_LGBM: phage virion proteins classification with UniRep features and lightGBM model, Comput. Math. Methods Med. 2022 (2022).

[22] B. Yang, W. Bao, J. Wang, Active disease-related compound identification based on capsule network, Brief. Bioinform. 23 (1) (2022) bbab462.

[23] V. Mazzia, F. Salvetti, M. Chiaberge, Efficient-capsnet: Capsule network with self-attention routing, Sci. Rep. 11 (1) (2021) 14634.

[24] W. Bao, B. Yang, B. Chen, 2-hydr_Ensemble: lysine 2-hydroxyisobutyrylation identification with ensemble method, Chemometr. Intell. Lab. Syst. 215 (2021) 104351.

[25] S. Opricovic, Programski paket VIKOR za visekriterijumsko kompromisno rangiranje, in: 17th International Symposium on Operational Research SYM-OP-IS, 1990.

[26] A. Hashemi, M.B. Dowlatshahi, H. Nezamabadi-pour, Ensemble of feature selection algorithms: a multi-criteria decision-making approach, Int. J. Mach. Learn. Cybern. 13 (1) (2022) 49–69.

[27] G. Yao, X. Hu, G. Wang, A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain, Expert Syst. Appl. 200 (2022) 117002.

[28] M. Rashid, J. Kamruzzaman, T. Imam, S. Wibowo, S. Gordon, A tree-based stacking ensemble technique with feature selection for network intrusion detection, Appl. Intell. 52 (9) (2022) 9768–9781.

[29] R. Tibshirani, The lasso method for variable selection in the Cox model, Stat. Med. 16 (4) (1997) 385–395.

[30] J.-X. Qian, M. Yu, Z. Sun, A.-M. Jiang, B. Long, A 17-gene expression-based prognostic signature associated with the prognosis of patients with breast cancer: A STROBE-compliant study, Medicine 99 (15) (2020).

[31] M. Zhang, K. Zhu, H. Pu, Z. Wang, H. Zhao, J. Zhang, Y. Wang, An immune-related signature predicts survival in patients with lung adenocarcinoma, Front. Oncol. 9 (2019) 1314.

[32] D. Kumar, B. Klefsjö, Proportional hazards model: a review, Reliab. Eng. Syst. Saf. 44 (2) (1994) 177–188.

[33] J.E. Herndon, S. Fleishman, A.B. Kornblith, M. Kosty, M.R. Green, J. Holland, Is quality of life predictive of the survival of patients with advanced nonsmall cell lung carcinoma? Cancer: Interdiscip. Int. J. Am. Cancer Soc. 85 (2) (1999) 333–340.

[34] K.-S. Wang, Y. Liu, S. Gong, C. Xu, X. Xie, L. Wang, X. Luo, Bayesian cox proportional hazards model in survival analysis of HACE1 gene with age at onset of alzheimer's disease, Int. J. Clin. Biostat. Biometr. 3 (1) (2017).

[35] L.-J. Wei, The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis, Stat. Med. 11 (14–15) (1992) 1871–1879.

[36] R. Saikia, M.P. Barman, A review on accelerated failure time models, Int. J. Stat. Syst. 12 (2) (2017) 311–322.

[37] G.K. Vishwakarma, P. Kumari, A. Bhattacharjee, Thresholding of prominent biomarkers of breast cancer on overall survival using classification and regression tree, Cancer Biomark.: Section A Dis. Markers (2021).

[38] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B Stat. Methodol. 34 (2) (1972) 187–202.

[39] A.P. Atanu Bhattacharjee, SurviMChd: High dimensional survival data analysis with Markov chain Monte Carlo, 2021, URL https://rdrr.io/cran/SurviMChd/.

[40] A. Bhattacharjee, P.K. Gajendra Kumar Vishwakarma, Afthd: Accelerated failure time for high dimensional data with MCMC, 2021.

[41] S. Narrandes, W. Xu, Gene expression detection assay for cancer clinical use, J. Cancer (2018).

[42] A.Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, ICML '04, Association for Computing Machinery, New York, NY, USA, 2004.

[43] D. Ghosh, A.M. Chinnaiyan, Classification and selection of biomarkers in genomic data using LASSO, J. Biomed. Biotechnol. 2005 (2) (2005) 147.

[44] M.M. Vasquez, C. Hu, D.J. Roe, Z. Chen, M. Halonen, S. Guerra, Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application, BMC Med. Res. Methodol. 16 (1) (2016) 1–19.

[45] H. Kim, M. Bredel, Feature selection and survival modeling in the cancer genome atlas, Int. J. Nanomedicine 8 (sup1) (2013) 57–62.

[46] A. Bhattacharjee, J. Dey, P. Kumari, A combined iterative sure independence screening and cox proportional hazard model for extracting and analyzing prognostic biomarkers of adenocarcinoma lung cancer, Healthcare Anal. 2 (2022) 100108, URL https://www.sciencedirect.com/science/article/pii/S277244252200048X.

[47] Y.-W. Wan, J. Nagorski, G.I. Allen, Z. Li, Z. Liu, Identifying cancer biomarkers through a network regularized cox model, in: 2013 IEEE International Workshop on Genomic Signal Processing and Statistics, IEEE, 2013, pp. 36–39.