Integrated Cox model for survival analysis and biomarker discovery with a feature ranking technique based on z-score transformation in non-small cell lung cancer patients.

Anidha M1, Premalatha K2

1Department of Information and Communication Engineering, Anna University, Chennai, India
2Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India

Abstract

Purpose: The purpose of this study was to identify signature biomarker miRNAs that are highly correlated with survival time and other clinical phenotypes. We have done the survival analysis with Cox proportional hazards model and presented a novel technique to identify signature biomarkers.

Materials and methods: The data set from Yu et al. consists of microarray miRNA expression profiles of 112 non-small cell lung cancer (NSCLC) patients. There are 55 adenocarcinoma, 50 squamous cell carcinoma and 7 large cell carcinoma patients. Among them 24 are females and 88 are males, 38 dead and 74 alive. Also, it has stage-I-47, stage-II-28 and stage-III-37 patients.

Results: This leads to the efficient calculation of risk score and the patients who are designated as high-risk patients. It deals with a novel biomarker selection/ranking technique zsTrans, through which the highly discriminant biomarkers are identified for various histological subtypes such as adenocarcinoma, and stages such as stage-I, stage-II and stage-III of non-small cell lung cancer (NSCLC) patients.

Conclusion: The identified biomarker miRNAs play vital role in the diagnosis and prognosis of NSCLC patients.

Keywords: Survival analysis, Cox proportional hazard model, miRNA, NSCLC, Biomarkers.

Accepted on January 18, 2016

Introduction

Non-small-cell lung cancer (NSCLC) causes more deaths than any other type of cancer worldwide. The recurrence rate among patients with early-stage NSCLC is about 40% within 5 years even after receiving combined surgical and chemotherapies [1]. The current tumour progression classification system is not an adequate measurement for treatment prognosis [1]. MicroRNAs are a new class of small non-protein-coding RNAs that function in endogenous negative gene-regulation and tumorigenesis. This effort may have clinical implication in the molecular-pathogenesis of cancer, development of new targeted-therapy or selection of high-risk cancer patients for adjuvant chemotherapy [1]. In spite of the difficulty in class prediction, there is an explosion of interest in biomarker research with the goal of incorporating biomarkers into drug development and leading to personalized medicine [2]. The potential utility of biomarkers is used for patient selection. By selecting patients based on their biomarker profiles, there is a hope to enrich the pool of patients who have a greater probability of response to alternative treatment plans. If successful, this approach could lead to cheaper and faster clinical trials than the conventional ones [2].

Researchers were able to identify sets of genes whose expressions are highly related to drug response of a set of compounds, which will eventually help development of new drugs. Correlating gene expression profiles obtained from tumor samples prior to treatment with the time to cancer relapse or death due to cancer can be very important in clinical practice [3]. The supervised approach is used to analyse the microarray data to identify biomarker miRNAs whose expression values are highly correlated with clinical or survival data which in turn used to predict the survival period of the patients. Li et al. stated that they have used kernel Cox model and negative partial likelihood as a loss function [3]. This method generalized the idea of support vector machines for binary or multi-categorical data to censored survival data. The model automatically searched for the genes whose expression levels are related to survival phenotypes and identifies the optimal combination of the gene expression data in predicting the risk of cancer recurrence or death [3]. Lossos et al. estimated the survival curve using Kaplan-Meier method and compared them using log-rank test [4]. Univariate cox model was constructed to identify biomarkers whose z-values greater than 1.5 or less than -1.5 are analysed. They have selected a group of 36 genes in this study [4]. Zhang et al. designed simulation techniques of microarray data with biological diversity related to treatment response or survival [2]. They developed Umpire R Package to simulate the connections.
Materials and Methods

Data set

The Data Set from Sung-Liang Yu et al., [7] consists of microarray miRNA expression profiles of 112 NSCLC patients. There are 55 Adenocarcinoma, 50 Squamous cell carcinoma and 7 large cell carcinoma patients. Among them 24 are females and 88 are males, 38 dead and 74 alive. Also it has Stage-I-47, Stage-II-28 and Stage-III-37 patients (Table 1).

Table 1. Description of data set.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Stage</th>
<th>Survival status</th>
<th>Histological type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>I</td>
<td>47</td>
<td>AD</td>
</tr>
<tr>
<td>Female</td>
<td>II</td>
<td>28</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>28</td>
<td>SQ</td>
</tr>
<tr>
<td></td>
<td>Dead</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Alive</td>
<td>38</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>AD</td>
<td>74</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>SQ</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Survival analysis

Survival analysis is a statistical method of modelling time to event (survival) data to analyse clinical outcome of patients. The event may be death, disease progression or recurrence of a tumor. It deals with the following

1. Rate of Survival at time period t.
2. The rate of failure or death at time t.
3. Effect of survival probability when there are changes in characteristics/circumstances.

Censoring, survival function and hazard are the common terms used in survival analysis. If a subject does not experience the event or outcome that is mentioned in the survival analysis within the study period is known as censoring. Right censoring refers to the end of the observation period of the subject before the event occurs. The event doesn’t occur beyond the study period; while left censoring refers to the event occurs before study period ends. Survival function or survival probability maps the event which is mentioned in the survival analysis with time and it captures the probability of the survival of the subject beyond the specified time period.

Let T be a continuous random variable with Cumulative Distribution Function (CDF) F (t) then

\[ S(t) = \int_0^\infty f(u)du = 1 - F(t) \rightarrow (1) \]

The Hazard which is known as instantaneous failure rate, denoted by \( h(t) \) or \( \lambda(t) \) is the probability of the event occurs at time t for an individual. In contrast to the survivor function, which focuses on not having an event, the hazard function focuses on the event occurring. It provides insight into the conditional failure rates and provides a vehicle for specifying a survival model. In summary, the hazard relates to the instantaneous (current) event rate, while survival reflects the cumulative non-occurrence [8].

\[ H(t) = \int_0^t f(u)/(S(u)) \rightarrow (2) \]

Where \( f(t) \) failure density functions or time to failure distribution.

Techniques of analysing survival data

These are the methods of estimating survival function and hazard function

Non-parametric methods:
1. Kaplan-Meier
2. Life table
3. Nelson-Aalen or Fleming-Harrington (via estimating the cumulative hazard)

Semiparametric method-Cox proportional hazards model

Parametric method Accelerated Failure Time (AFT) model

Kaplan-Meier (K-M): Survival probability is estimated non-parametrically from survival data which works with conditional and cumulative probabilities. The probabilities of surviving from one interval to the next is multiplied together to give the cumulative survival probability.

Conditional probability \( p_i=r_i-d_i/r_i \rightarrow (3) \)

Where \( r_i \) is the number of alive persons at time \( t_i \) and \( d_i \) is the number of failures (dead) persons at the same time interval.

Survival at any time \( t \) is calculated as a cumulative probability by multiplying the conditional probabilities of surviving each previous time interval.

\[ S(t)=\prod_{i=1}^{t}(1-d_i/r_i) \rightarrow (4) \]

Life table: If there are no records available about the exact time of the events occur then the time intervals are taken into account with the basic assumption of uniformly distributed censorings. The average number of subjects at risk at time interval t is calculated as
**Integrated Cox model for survival analysis and biomarker discovery with a feature ranking technique based on z-score transformation in non-small cell lung cancer patients**

\[ n_j = n_j - C_j/2 \rightarrow (5) \]

\[ n_j^c \] is the number of subjects alive at time interval \( j \) and \( c_j \) is number of censored observations. The survival probability is calculated as

\[ S(t) = \prod_{j \leq t} (1 - d_j/n_j^c) \rightarrow (6) \]

**Nelson-Aalen estimator:** This method is non-parametric hazard rate estimator when the data is censored or incomplete. The survival probability is calculated as

\[ S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \rightarrow (7) \]

\( d_i \) is the number of events at time \( t \) and \( n_i \) is the number of individuals (subjects) at the same time period.

**Semi-parametric: Cox proportional hazards model:** The K-M estimator is used for univariate analysis for time-to-event outcome measure of survival analysis whereas the Cox model is used for multivariate survival analysis. The general Cox model is known as

\[ H(t) = h_0(t) \times e^{(b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)} \rightarrow (8) \]

The hazard function has two terms. \( h_0(t) \) is known as baseline hazard which depends on time not covariates. \( e^{(b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)} \) depends on covariates not time. If we are interested in the effects of covariates then there is no need to include the term baseline hazard \( h_0(t) \). The Cox model is thus known as semi-parametric method.

The hazard ratio for two observations is described as follows,

\[ \frac{h(x_1, t)}{h(x_2, t)} = \frac{e^{b_1 x_1}}{e^{b_2 x_2}} = e^{[b_1 (x_1 - x_2)]} \rightarrow (9) \]

Hazards are proportional to each other. The hazard for the observation with covariate \( x_1 \) is \( e^{b_1 (x_1 - x_2)} \) times that of hazard ratio of observation with covariate \( x_2 \). The term \( e^{b_1 (x_1 - x_2)} \) is hazard ratio comparing covariate \( x_1 \) to covariate \( x_2 \). The hazard ratio greater than 1 indicates that if the value of covariate of \( p \)th observation increases, the event hazard increases and thus the rate of survival of that particular observation decreases. A hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival [9]. This is known as Assumption of proportionality related to Cox model.

**Survival analysis using supervised principle components**

The technique behind the SPC is rather than performing principle component analysis with the entire set of features, it is performed with the set of features which has strong correlation with survival outcome. First principle component is computed using this strongly correlated feature set [10].

**Algorithm for computing supervised principle components:**

1. Univariate regression coefficients are computed for each feature.
2. Data matrix is reduced with features whose regression coefficients are greater than a threshold.
3. Compute the first principle components for this reduced data matrix.
4. Use these principle components in the regression model to predict the outcome.

Comparison of Survival Analysis Techniques is shown in Table 2.

**Experimental Design**

We propose integrated Cox model to perform the risk analysis to compute the risk scores of the samples. Based on the risk scores the samples are classified into high risk group and low risk group. Further analysis of identifying dominating biomarkers in each stage and histological subgroups are carried out with high risk group samples.

Conditional hazard vector \( V \) is defined as follows,

\[ V = b_1, b_2, ..., b_n \rightarrow (10) \]

where \( b_i \) is the coefficients of features of univariate general cox model whose hazard rates are \( \geq 1 \) and \( n \) is number of features which satisfy the above criteria.

\[ T = t_1, t_2, ..., t_n \rightarrow (11) \]

where \( t_1, t_2, ..., t_n \) are the coefficients of covariates from the univariate general cox model.

Then we perform the Integrated Cox model with the following

\[ h(t) = h_0(t) e^{(bX+T)} \rightarrow (12) \]

where \( X = x_{11}, x_{12}, ..., x_{1m} \) expression values of features whose univariate cox coefficients are \( \geq 1 \).

We can rewrite the Equation 12 in terms of risk score function as

\[ h(t) = h_0(t) e^{(bX)} \rightarrow (13) \]

The risk score function is written as

\[ f(X) = \sum_{j=1}^{p} \left\{b_j x_j\right\} + T \rightarrow (14) \]
zsTrans feature ranking technique

We propose a novel feature ranking technique as a biomarker identification method to identify differentially co-expressed, dominating features related to each Stage of the NSCLC such as Stage-I, II, III and different histological subgroups such as adenocarcinoma, squamous cell carcinoma and large cell carcinoma.

General Cox model for multivariate analysis with interaction is,

\[ H(t) = h_0(t)e^{\beta X_t} \]  \( \rightarrow \) (15)

Where \( X_t = \{x_1, x_2, x_{12}\} \) and \( \beta = \{b_1, b_2, b_{12}\} \)

From Equation 15 we get z-score values of all the features and it is defined as

\[ Z = z_{ji} \]  \( \rightarrow \) (16)

where \( i = 1, \ldots, n; n = \text{number of features and } j = 1, \ldots, p; p = \text{number of samples} \). Then we do z-Score normalization of expression values for all the features of samples which satisfy the ranking criteria which is denoted as \( p_1 \) and the feature ranking vector \( V \) is defined as

\[ V = v_1, v_2, \ldots, v_{p1} \]  \( \rightarrow \) (17)

Where \( v_i \) is defined as

\[ v_i = \frac{1}{p_1} \sum_{j=1}^{p1} \{z_{ji} - \bar{z}_{ji}\} \]  \( \rightarrow \) (18)

All the features are ranked according to their \( v_i \) values. We used this technique to identify top ranked features as a dominating biomarkers for each histological subgroups and Stage-I, II and III for high risk samples which are selected from the entire dataset using the risk score.

Results and Discussions

Univariate Cox analysis was performed on the 157 miRNAs to identify features associated with survival in the training set of 70 samples. There are 6 miRNAs identified which are strongly correlated with patient’s survival. P-values which are \( \leq 0.05 \) for Death as a result of univariate Cox regression analysis are used to identify the signature miRNAs. Table 3 shows the identified miRNAs and their HR, Coeff and p values.

We then performed the supervised principle components analysis and confirmed the selection of miRNA signature model for risk score analysis with the first principle components. Table 4 shows the feature scores of first principle components and Figure 1 shows the selection of threshold which fits the model with the outcome using the likelihood ratio statistics, Figure 2 shows the survival probability of high risk and low risk groups, Figure 3 shows the reduced set of features based on the feature scores computed using supervised principle components.

The Figures 1-3 shown here are the results of the supervised principle component analysis.

Table 3. Signature miRNAs.

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Hazard ratio</th>
<th>Coeff-b value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR137</td>
<td>1.164</td>
<td>0.15</td>
<td>0.011</td>
</tr>
<tr>
<td>miR372</td>
<td>1.364</td>
<td>0.31</td>
<td>0.022</td>
</tr>
<tr>
<td>miR182</td>
<td>1.323</td>
<td>0.27</td>
<td>0.037</td>
</tr>
<tr>
<td>miR147</td>
<td>1.27</td>
<td>0.23</td>
<td>0.056</td>
</tr>
<tr>
<td>miR221</td>
<td>0.874</td>
<td>-0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>Let7a</td>
<td>0.867</td>
<td>-0.14</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Table 4. First principle components.

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Feature score</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR 137</td>
<td>1.52</td>
</tr>
<tr>
<td>miR 372</td>
<td>1.03</td>
</tr>
<tr>
<td>miR 182</td>
<td>1.04</td>
</tr>
<tr>
<td>miR 147</td>
<td>0.92</td>
</tr>
<tr>
<td>miR 221</td>
<td>-1.65</td>
</tr>
<tr>
<td>Let 7a</td>
<td>-1.49</td>
</tr>
</tbody>
</table>

Figure 1. Likelihood ratio statistic for selecting threshold to fit the model with the outcome.

Figure 2. Survival probability of high risk and low risk groups (Red line=Low risk group; Green line=High risk group).
We have performed the univariate cox analysis for other covariates such as gender, stage, histological subtype and the quantiles of age. The resultant coefficients are considered appropriately while computing risk score for each sample. Based on the risk scores estimated, the samples are partitioned into high risk groups and low risk groups using median value as a threshold. Figures 4 and 5 Show the Kaplan-Meier survival curves for the two risk groups in training set and test set.

We have analysed that the 6 miRNA signature model is significantly associated with the covariates using the multivariate cox proportional hazard analysis. We have found that Stage II and III are strongly associated with the risk and survival of the patients. Histological subtype adenocarcinoma is found to be more risky than the squamous cell carcinoma whose hazard ratio is 2.64 and p-value is 0.03. When the covariate gender is analysed, we get 1.18 as HR for female and 0.556 for male patients and age is less significant than the other covariates in the model (Figures 6-10 and Tables 5 and 6).

**Table 5. Multivariate Cox analysis.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hazard ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage II</td>
<td>2.42</td>
<td>0.08</td>
</tr>
<tr>
<td>Stage III</td>
<td>5.49</td>
<td>0.00095</td>
</tr>
<tr>
<td>Age</td>
<td>1.01</td>
<td>0.19</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>2.64</td>
<td>0.03</td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>0.377</td>
<td>0.01</td>
</tr>
<tr>
<td>Gender-Male</td>
<td>0.556</td>
<td>0.03</td>
</tr>
<tr>
<td>Gender-Female</td>
<td>1.18</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**Table 6. Results of ANOVA.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Loglik</th>
<th>Chisq</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-144.99</td>
<td>3.28</td>
<td>0.06</td>
</tr>
<tr>
<td>Stage</td>
<td>-132.72</td>
<td>24.53</td>
<td>0.00046</td>
</tr>
<tr>
<td>Cell type</td>
<td>-146.64</td>
<td>7.39</td>
<td>0.02</td>
</tr>
<tr>
<td>Gender</td>
<td>-132.06</td>
<td>1.32</td>
<td>0.25</td>
</tr>
</tbody>
</table>

To test the statistical significance of the K-M curves for two or more groups, we have done log-rank test. The log-rank statistic
makes use of observed vs. expected values over categories of outcome.

**Table 7. Stages of NSCLC.**

<table>
<thead>
<tr>
<th>Stages</th>
<th>Events observed</th>
<th>Events expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I</td>
<td>8</td>
<td>18.5</td>
</tr>
<tr>
<td>Stage II</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Stage III</td>
<td>22</td>
<td>10.5</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>38</td>
</tr>
</tbody>
</table>

The log-rank statistic for Stage is 18.91 and the p-value is 0.00007825 which indicates that the null-hypothesis is rejected. This denotes that the risk and survival rate are not same with respect to the stages of the NSCLC patients (Table 7).

**Table 8. Histological subtypes.**

<table>
<thead>
<tr>
<th>Histological subtype</th>
<th>Events observed</th>
<th>Events expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenocarcinoma</td>
<td>23</td>
<td>16.83</td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>12</td>
<td>18.58</td>
</tr>
<tr>
<td>Large cell carcinoma</td>
<td>3</td>
<td>2.58</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>38</td>
</tr>
</tbody>
</table>

The log-rank statistic for Histological subtype is 4.69 and the p-value is 0.09572 which indicates that the null-hypothesis is rejected. This denotes that the risk and survival rate are slightly different with respect to the histological subtypes of the NSCLC patients (Table 8).

**Figure 8. K-M estimate of overall survival of stage -III (n=22).**

**Figure 9. K-M estimate of overall survival of adenocarcinoma (n=33).**

**Figure 10. K-M estimate of overall survival of squamous cell carcinoma (n=31).**

**Table 9. Gender.**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Events observed</th>
<th>Events expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>30</td>
<td>31.07</td>
</tr>
<tr>
<td>Female</td>
<td>8</td>
<td>6.93</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>38</td>
</tr>
</tbody>
</table>

The log-rank statistic for gender is 0.2 and the p-value is 0.6511 which indicates that the null-hypothesis is not rejected. This denotes that the risk and survival rate are same with respect to the sex of the NSCLC patients (Table 9).

Multivariate Cox analysis with interaction of covariates was performed and the z-scores are used for the ranking of the features in each stage and histological subtype of high risk group patients (Table 5). We have done z-score normalization of expression values based on Equation 18 and we obtained scores for all the features and the features are ranked accordingly.

**Table 10. Biomarker miRNAs identified related to histological subtype using the ranking technique.**

<table>
<thead>
<tr>
<th>Adenocarcinoma</th>
<th>Up-regulated miRNAs</th>
<th>Down-regulated miRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feature name</td>
<td>Z-score value</td>
</tr>
<tr>
<td></td>
<td>miR-200a</td>
<td>0.313</td>
</tr>
<tr>
<td></td>
<td>miR-135b</td>
<td>0.023</td>
</tr>
</tbody>
</table>
The microRNA-200 (miR-200) family is a powerful regulator of the Epithelial-Mesenchymal Transition (EMT) process, which is essential in tumor metastasis [11]. The identification of miR-200 family includes miR-200a, miR-200b and miR-200c are functionally involved in canonical pathways of immune response, molecular mechanisms of cancer, metastasis signalling, cell-cell communication, proliferation and DNA repair in Ingenuity Pathway Analysis (IPA) [11]. These provide a potential basis for innovative therapeutic approaches for the treatment of this deadly disease. Low miR-100 expression was found to be closely correlated with higher clinical stage, advanced tumor classification and lymph node metastasis of patients. The overall survival of NSCLC patients with low miR-100 was significantly lower than that of those patients with high miR-100 [12]. The overexpression of miR-210 is highly associated with an improved disease specific survival and it is a candidate marker for prognostic stratification in NSCLC [13]. The high co-expression of miR-27a results in increased risk of cancer and poor prognosis for Non-Small Cell Lung Cancer (NSCLC) [14]. The miR-99a was downregulated in NSCLC tissues and suppressed tumor metastasis of NSCLC cells. Down-regulation of miR-99a is significantly associated with last-stage and tumor metastasis in NSCLC patients [15]. Platinum-based chemotherapies have long been used as a standard treatment in non-small cell lung cancer. However, cisplatin resistance is a major problem that restricts the use of cisplatin. Deregulated cell death mechanisms including apoptosis and autophagy could be responsible for the development of cisplatin resistance and miRNAs are the key regulators of these mechanisms. The miR-15a is an important key regulator which enhances the anticancer effects of cisplatin in the resistant non-small cell lung cancer cells [16]. MiR-195 suppresses tumor growth and is associated with better survival outcomes in Non-Small Cell Lung Cancer (NSCLC). It is showed that high miR-195 plasma levels associated with favourable overall survival of non-smoking women with lung adenocarcinoma [17]. The high miR-146 expression level longer progression-free survival (25.6 weeks in miR-146a high patients vs. 4.8 weeks in miR-146a low patients, P<0.05). miR-146a is therefore a strong candidate prognostic biomarker in NSCLC. Thus inducing miR-146a might be a therapeutic strategy for NSCLC [18].

### Table 11. Biomarker miRNAs identified related to stage using the ranking technique.

<table>
<thead>
<tr>
<th>Stage I</th>
<th>Up-regulated miRNAs</th>
<th>Down-regulated miRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature name</td>
<td>Z-score value</td>
<td>HR</td>
</tr>
<tr>
<td>miR-130a</td>
<td>0.076</td>
<td>1.116</td>
</tr>
<tr>
<td>miR-15b</td>
<td>0.444</td>
<td>1.204</td>
</tr>
</tbody>
</table>
miR-218  -1.309  0.697  1.3  miR-325  -0.146  0.975  372.72
miR-20   0.837  1.363  1.5  miR-183  -0.106  0.985  348.63
miR-27a  0.372  1.107  4.53 miR-134  -0.099  0.981  289.24
miR-182  -0.733  0.876  4.7  miR-302a  -0.222  0.962  242.46

miR-218  -2.153  0.57   0.26  miR-325  -0.315  0.96   132.24

miR-222  -2.01   0.47   0.02  miR-104  -0.023  0.998  420.13
miR-221  -2.396  0.792  0.03  miR-193-a*  -0.1518 0.976  267.86
miR-342  0.419  1.208  0.03  miR-299  -0.0912 0.981  148.42
miR-98   -2.975  0.65   0.08  miR-371  -0.3161 0.95   139.32
miR-199-s 1.154  1.629  0.11  miR-183  -0.547  0.917  137.73
miR-218  -2.153  0.57   0.26  miR-325  -0.315  0.96   132.24

The miR-142-5p is identified as a potential biomarker for diagnosis and treatment of squamous lung cancer. In NSCLC the miR-155 is identified as an oncogene and is associated with poor prognosis and the miR-335 is strongly associated with small cell lung cancer metastases [20]. It is demonstrated that the overexpression of miR-221 is correlated with proliferation rate and cell cycle phase distribution in Lung cancer and other cancers such as breast cancer. The overexpression of miR-301 is highly associated with poorly differentiated lymph node metastasis and advanced tumor node metastasis and it is unfavourable prognostic factor for disease-free survival [21]. The aberrant expression of hsa-miR27a results in increased risk of cancer and poor prognosis of NSCLC and it is proved that miR-27a could suppress RKIP expression and in turn contribute to chemoresistance of lung adenocarcinoma cells to cisplatin [22,23]. The miR-197 is a novel biomarker related to chemoresistance and survival of NSCLC. The miR-197/CKS1B/STAT3 axis has the ability to promote cancer progression in chemoresistant NSCLC. It is suggested that miR-197 replacement therapy may be an effective treatment for lung cancer patients with chemoresistance, particularly in PD-L1-positive patients [23]. The administration of antisense of miR-183 significantly increased FoxO1 levels in NSCLC resulting in a significant decrease in NSCLC growth (Table 11). It is demonstrated that miR-183/FoxO1 axis may be a novel therapeutic target for regulating the growth of NSCLC [24].

**Conclusion**

Lung cancer is the remarkable cause of the cancer related deaths and approximately 70% of the newly diagnosed patients are in advanced or metastatic state. The miRNAs play critical role in cell growth, proliferation, differentiation and development of various solid and haematological malignancies. Recently, miRNAs have emerged in NSCLC as both diagnostic and prognostic biomarkers. The integrated Cox model suggested here is proved as a novel technique which differentiates High risk and low risk patients using the effective computation of risk scores. A novel feature ranking technique based on z-score transformation is used to rank and identify novel biomarkers in different stages and histological subtypes of high risk NSCLC patients.

**Funding**

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.
Conflicts of Interest

The authors state that there are no conflicts of interest regarding the publication of this article.

References


*Correspondence to*

Anidha M
Department of Information and Communication Engineering
Anna University
Chennai
India