Identifying Highly Predictive Pretreatment MicroRNA Signature For Cardiac Events In Lung Cancer Patients After Treatment

5724

March 9, 2017

Abstract

In this study, we performed survival analysis on 63 patients with non-small-cell lung cancer after conformal radiotherapy, in order to find a highly predictive pattern of microRNA signature that could predict the risk of having cardiac events greater than grade 2 after treatment. By building a Cox proportional hazards model with elastic net penalty and stepwise regression, we selected 15 microRNAs that could predict cardiac events better than just using baseline variables such as Framingham Risk Score.

Introduction

MicroRNAs (miRNAs) are small endogenous ~22 nucleotides RNAs that regulate the expression of complementary messenger RNAs (Victor Ambros 2004)(David P Bartel 2004). miRNA regulation can be involved in the cell developmental fate decisions, but can also have more subtle roles in buffering stochastic fluctuations in gene expression (K Musilova 2015). Therefore, microRNA signature can serve as promising biomarkers for early detection and prognosis of human cancer and many other human diseases. In this study, our objective is to determine whether microRNA signature can predict the risk of cardiac event at certain time points after the conformal radiotherapy for patients with non-small-cell lung cancer, and if so, what the predictive pattern of microRNA is. Specifically, we are looking at whether the microRNA signature can improve the accuracy of prediction of the risk of having cardiac events, which is currently based on baseline information at treatment, such as age, gender, smoking status, cardiac event history, comorbid status etc.

Methods

Clinical and biomarker data were collected from 63 patients with non-small-cell lung cancer, who were treated at University of Michigan Hospital and Veterans Hospital in Ann Arbor. The treatment start date ranged from 2004 to 2011, and patients were followed up after treatment until they died or left the study.

One of the main measures of the risk of future cardiac event provided in the data was the Framingham Coronary Heart Disease Risk Score, which was a gender-specific algorithm used to estimate the 10-year cardiovascular risk of an individual. This score was only defined for patients who did not have baseline cardiac disease, and in the data they were represented as "NA". Therefore, we decided to first categorize the Framingham Score based on gender and risk level, and then assign a score to those with cardiac event history.

After data preprocessing, we performed survival analysis by building Cox proportional hazards models. Our methods took the following steps:

1. Fit a Cox proportional hazards model with only selected baseline predictors (without the microRNAs). We named this model as Model 1, and it would serve as a benchmark for model comparison.
2. Fit a Cox proportional hazards model with only the 61 microRNAs, and used elastic-net penalty and 10-fold cross validation to find the most representative subset of microRNAs.
3. Checked if multicollinearity was an issue by constructing a correlation matrix, and comparing variance inflation factors (VIF), and removed predictors with high VIFs if there were any. We named the model we had after this step as Model 2.
4. Repeated Step 2 with the subset of microRNAs we currently had until no more microRNAs were screened. (Model 3)
5. Performed a stepwise regression on Model 3 to further reduce the number of predictors. (Model 4)

For each step in step 2-5, we added the subsets of microRNAs as predictors into our benchmark model (Model 1), and recorded the regression results, AIC, BIC, etc., in order to choose our final model.

Results

The patients in the data were treated in two sites: 23 (36.51%) at the University Hospital, and 40 (63.49%) at the Veterans Hospital. The mean age of the patients at baseline was 66.4, ranged from 45.30 to 84.60. The

Table 1: Framingham Score Categorization

Category	Risk Level	Men	Women
3	Extremely High Risk	Pre-existing Cardiac Disease	Pre-existing Cardiac Disease
2	High Risk	F-score >= 17	F-score >= 25
1	Intermediate Risk	12 <= F-score < 17	20 <= F-score < 25
0	Low Risk	F-score < 12	F-score < 20

majority of the patients were white male (48 (76.19%) male and 59 (93.65%) white). The mean follow-up time was 30.62 months, ranged from 2.70 months to 110.10 months, with 49 (77.78%) died during follow-up. The outcome variables we were interested in were the grade two cardiac event status, which indicates whether the patients had a cardiac event that had a grade of at least two, and the time in months from the start of the treatment to the first grade two events. In the data, 28 (44,44%) of the patients had a grade two cardiac events.

Based on the scoring system available online,[1] we categorized the Framingham Score as in Table 1. Among the 63 patients, 19 were categorized as extremely high risk, 17 were categorized as high risk, 8 were categorized as intermediate risk, and 19 were categorized as low risk.

Figure 1 showed the Kaplan-Meier Curve for the four Framingham categories. We could observe that patients in low risk category had a higher survival probability than patients in the other three categories. However, the differences among the other three categories were not obvious. This told us that Framingham category might be a good predictor of cardiac events, but there was still space for improvements.

We first fit a Cox model with T-stage, heart mean dose, and Framingham category:

$$\lambda_i(t) = \lambda_0(t) exp(\beta_1 * T.Stage + \beta_2 * Heart\_Meandose + \beta_3 * F\_Category)$$

The outcome variable was the risk of getting grade 2+ cardiac event at time t. T stage described the size of the original (primary) tumor and whether it had invaded nearby tissue. The higher the value, the larger the size of the tumor would be. We predicted that larger tumor would lead to higher risk of cardiac events. "Heart mean dose" was the mean dose of radiation to the heart in gray, and we predicted that the higher the dose, the higher the risk of cardiac events. "F_Category" was the Framingham category that we created.

The result confirmed our predictions, with Heart Meandose and F category significant at 95% level, and T-Stage significant at 90% level. All three coefficients were positive. The result told us that patients with

[1] https://en.wikipedia.org/wiki/Framingham_Risk_Score

Table 2: Model 1 Confusion Matrix

	Event	No Event
High Risk	25.40%	23.81%
Low Risk	6.35%	44.44%

one level higher in Framingham category were 59.17% more risky in getting grade 2 cardiac event, holding all other covariates constant. Figure 2 was a residual plot in which x-axis was the deviance residual, and y axis was the id of the patients. The plot showed that the residuals were randomly scattered around 0, which demonstrated that the model fit the data well.

We used the prediction function to generate predicted values using the Cox Model developed. We then used the basehaz function to calculate the baseline hazard at different time points. Now we could calculate the risk of having grade 2 cardiac events at different time points. We categorized the subjects into two categories: high risk and low risk, by the median of the predictive value. Table 2 was a confusion matrix that showed the proportion of true positive, false positive, true negative and false negative of the prediction at 12 months. Based on the confusion matrix, the accuracy of the prediction was 25.40% + 44.44% = 69.84% (true positive rate + true negative rate).

Variable Screening – MicroRNAs

Next we fit a Cox PH Model with only the 61 microRNAs and tried to find a predictive pattern. To do the variable screening we needed to perform some types of model complexity regularization because we had more predictors than subjects otherwise. To avoid over-fitting, we performed 10-fold cross validation by using the cv.glmnet function, and ran the function 100 times to find the best lambda, which was a tuning parameter that controlled the overall strength of penalty. However, due to the randomness of cross validation, we still got different optimal lambdas every time. So we ran the function for multiple 100-cycles and we found two lambda values that came up most often. One was 0.2, and the other was 0.031. We then plugged the variables selected by these two lambda values back to our Model 1, and we found the model with the variables selected by lambda = 0.031 had the best BIC. Therefore, our next model would contain 22 microRNAs: NA05, NA06, NA07, NA12, NB04, NB07, NB08, NC04, ND03, ND11, NE01, NE03, NE04, NE06, NE10, NF01, NF03, NF09, NF10, NG06, NG10, NG11. By adding these 22 microRNAs to our model 1, our model 2 looked like this:

$$\lambda_i(t) = \lambda_0(t)exp(\beta_1 * T.Stage + \beta_2 * Hart_Meandose + \beta_3 * F_Category + \beta_4 * NA05 + ... + \beta_{25} * NG11)$$

The number of microRNAs in the model was still very high, which suggested that there might be some predictors that did not contribute to the accuracy of the predictive model or might in fact decrease the accuracy of the model. Also, simple models were easier to understand and explain. Therefore, the goal of our following work was trying to reduce the complexity of the model, and at the same time keeping or improving the accuracy of the model.

Multicollinearity

We could reduce the number of microRNAs in our model by dealing with multicollinearity, because multicollinearity could reduce the predictive power of our model if presented. We first plotted a correlation matrix for all the 22 microRNAs we had selected. We could identify the pairs of predictors that had the highest correlation and kept only one predictor for each of those pairs. However, just looking at correlations among pairs of predictors might not be enough, because it was possible that the pairwise correlations were small, and yet a linear dependence existed among three or even more variables. Therefore, we calculated the variance inflation factors (VIF), and recursively removed the predictors that had VIF greater than a threshold. A VIF greater than 10 was often regarded as indicating multicollinearity. If we set the threshold to 10, all of the 22 variables were kept and the maximum VIF was 7.66. If we set the threshold to 5, two variables were removed. However, when we fit the Cox model with the 20 variables left, the accuracy was still the same as it was with 22 variables, but the BIC increased. If we set the threshold to 2.5, 6 variables were removed, but the accuracy was lower, and BIC higher than with 22 variables. In the end, it seemed that multicollinearity was not an issue, and keeping all the 22 microRNAs in the model was the best choice.

Another run of glmnet

We could further reduce the size of the feature space by performing elastic net on the 22 selected variables until no variables were removed. If we perform elastic-net again by repeating the procedure described above on the 22 selected microRNAs, we got another 2 microRNAs removed: NA06 and NE04. And if we perform a third run of elastic-net on the 20 microRNAs, no more mircroRNAs were removed. Therefore, we had our Model 3 with 20 microRNAs:

$$\lambda_i(t) = \lambda_0(t)exp(\beta_1 * T.Stage + \beta_2 * Hart_Meandose + \beta_3 * F_Category + \beta_4 * NA05 + ... + \beta_{23} * NG11)$$

Table 3: Model 4 Confusion Matrix

	Event	No Event
High Risk	31.75%	17.46%
Low Risk	0.00%	60.79%

Table 4: Model Comparison

Model	# MicroRNAs	AIC	BIC	Concordance	Accuracy
1	0	185.6	189.3	0.75	69.84%
2	22	147.3	177.8	0.95	79.37%
3	20	146.7	174.7	0.94	82.54%
4	15	139.6	161.6	0.94	82.54%

Stepwise Regression and Model Selection

By performing bidirectional elimination on Model 3, we ended up with 15 microRNAs: NA05, NA07, NB04, NB07, NB08, ND03, ND11, NE01, NE03, NE06, NE10, NF01, NF03, NF09, NG06. This model (Model 4) gave the lowest AIC and BIC, and highest accuracy (same as Model 3). Therefore, we chose this as our final model. Table 3 was the confusion matrix for our final model. We could notice that the accuracy of our final increased significantly comparing with our base model.

In Table 4 we made a comparison of the four models we constructed at each step with respect to AIC, BIC, concordance and accuracy of prediction of the risk of getting grade 2+ cardiac event at 12 month. In Figure 3 and 4, we made a comparison of the KM plot between the final model and our Model 1. The patients were categorized into four risk levels based on their predicted risk at 12 months. We could observe that the different categories were mode separated in the final model, which suggested that our final model did a better job in prediction.

Conclusion

To conclude, pretreatment microRNA signature could help us better predict the risk of cardiac events in lung cancer patients after treatment. Our final selected predictive microRNA signature contained 15 microRNAs, and the model with the set of microRNAs included had better predictive power and fit the data better with respect to AIC and BIC. Since the majority of the subjects in the data were white male from Ann Arbor area, the generalizability of our final model might be limited. Nevertheless, similar methods could be applied in the future in attempting to increase predictive power of models under similar settings.

Appendix

R Code is attached.

Reference

David P Bartel. 2004. "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." Journal Article.

K Musilova, M Mraz. 2015. "MicroRNAs in B-Cell Lymphomas: How a Complex Biology Gets More Complex." Journal Article.

Victor Ambros. 2004. "The Functions of Animal MicroRNAs." Journal Article.