Design and implementation of cancer patient survival prediction model based on ensemble learning method

Bi Chuanmei¹, Fang Xiaonan^{1*}, Geng Zezheng¹, Dong Ling¹

¹SHANDONG MANAGEMENT UNIVERSITY, 250357, China

Abstract. To study the impact of genomics on cancer diseases, bioinformatics and integrated learning methods are used to conduct survival analysis on colon cancer and rectal cancer data in the cancer gene map database. Firstly, according to the significant expression and stability test, the long-chain non-coding RNA in the transcriptome that has a significant impact on clinical prognosis survival analysis was initially screened. Then use a random forest ensemble learning algorithm to train it to get a preliminary model. Finally, based on the optimized random survival forest model, the Cox regression model was once again integrated, and the risk values of the two were integrated. The RCCT (Random-Cox Combined to Survival) method was proposed to provide clinical decision-makers certain reference values.

1 Introduction

The survival analysis method refers to the method of solving specific problems related to survival, and it is an important theoretical basis for establishing a survival prediction model and analyzing survival prediction ability. Therefore, by studying a variety of survival analysis methods to improve the predictive ability of survival analysis, it has important research significance and value for many fields, especially clinical prognosis analysis¹. Through the predictive model, assisting doctors to make decisions about the situation, to a certain extent alleviate the huge psychological and physiological pain of patients.

The Ensemble Learning method is to obtain multiple learners² through learning training samples, combine them to obtain an aggregate, and make predictions as a whole. Integrated learning eliminates the instability of learning by a single learner, and combines the advantages of multiple learners, which makes the prediction effect better and improves the generalization ability.

Survival analysis is of great significance in medical prognosis, using clinical data to predict the survival of patients after a certain period of time. The main goal is to be able to find the factors that affect the survival of patients and to establish a survival prediction model through the found influencing features to assist medical decision-making. To study when an event occurs in the observed object, it can use some methods and techniques to conduct survival analysis based on the patient's real data. In recent years, in addition to the application of statistical methods, machine learning methods have also been gradually applied to the research of survival analysis³. The ensemble learning method also has certain applications for the clinical prognosis and survival

analysis of cancer. In this study, long non-coding RNA was extracted from the genome for research. It not only has regulatory functions but also participates in biological processes and pathways. The occurrence of various cancers is inseparable from it. Although there are more lncRNAs than mRNAs, the role played by most lncRNAs is still unknown. Therefore, the research on lncRNA has important practical significance.

The traditional survival analysis method is based on the principles of statistics to predict the duration of an event after it occurs. For example, the survival time of patients from the onset of the disease has been widely studied as early as the 17th century. It is based on the analysis of a certain risk function, proposed by Gompertz and Makeham⁴ in the 19th century, based on statistical methods such as parameterized models such as maximum likelihood estimation. Among them, the Cox model⁵, as the most classic prediction model, is still widely used. Its prediction target is a risk ratio, that is, the ratio of an individual's risk function to the group's risk function. Assuming that the model has a linear relationship with the predicted risk factor, this method is based on a linear relationship model with good results. But for data that is not linear, its prediction effect is not particularly ideal. Derived from the above-mentioned machine learning, an integrated learning method has emerged. Ishwaran et al.⁶ built a Random Survival Forest (RSF) model based on the random forest method and combined with the traditional survival analysis method, and conducted more targeted research on survival analysis. The traditional survival analysis has been extended to a certain extent, and its model does not need to meet certain premises and assumptions, which makes up for the weaknesses of traditional survival analysis methods. In 2019, Sang Haokai⁷ and others used LASSO, Cox, and RSF methods for four omics data sets and

* Corresponding author: Fang Xiaonan, franknan@126.com.

explored their respective mechanisms. However, the method of acquiring features does not make full use of prior knowledge. And this method has different results on different cancer types, which increases the complexity of the model.

This research is based on the existing survival analysis model, combined with machine learning, especially ensemble learning methods, using rectal cancer and colon cancer data sets^{8,9} to construct a survival prediction model. On this basis, the proportional hazard model and the random forest survival model were once again integrated, and the RCCT method was proposed, and a more significant lncRNA was found for survival prediction of clinical prognosis.

2 Data preprocessing and feature extraction

2.1 Data collection

The data set used is located on TCGA. This data platform is a joint collection of human cancer data sets by multiple research institutes, including genome, transcriptome, miRNA, methylation, and other data.

The lncRNA and clinical data in the transcriptomics of rectal cancer and colon cancer data are selected for research. Among them, there are a total of 167 samples of rectal cancer data; a total of 472 samples of colon cancer data.

2.2 Data processing

Firstly, the data is preprocessed. There are a large number of 0 values in the transcriptome, and the expression level is low. Add 1 to the whole to avoid the influence of 0. The clinical data are integrated into 3 columns, and finally, they are standardized to avoid large numbers that affect the training effect of the model.

Secondly, the data is merged, the gene annotation file is used to extract the lncRNA data in the transcriptome, and the patient number of the data is used to integrate the clinical data and the lncRNA data in the transcriptome. After integration, rectal cancer data has 158 samples with 5,034 features, and colon cancer data has 433 samples with 4,992 features. Finally, delete the duplicate rows. The data set is divided into training and testing samples at 1:1, and the distribution of survival status 0 and 1 is 1:1.

2.3 Feature screening

Preliminary single-factor and multi-factor survival analysis were performed on all characteristics, and the correlation between variables and variables and between variables and survival period was analyzed.

Then the stability test¹⁰ is performed on each factor and combined with the *p value*, the significant factors are initially screened out. Among them, the *p value* of rectal cancer and colon cancer are all less than 0.05 during the univariate analyses. Then combined with the results of

multi-factor analysis, the final results of preliminary screening of lncRNA data were 6 significant factors for rectal cancer (Table 1) and 14 significant factors for colon cancer (Table 2).

The transcription type antisense means antisense, which is produced in the antisense strand of the coding gene. Sense intronic represents the intron lncRNA, which is mainly produced in the intron region of the coding gene. Processed transcript means re-transcription. lincRNA refers to the lncRNA in the intergenic region, which is produced in the middle region of the two coding genes. READ represents the rectal cancer data set, and COAD represents the colon cancer data set.

Table 1. Significant genes in the READ data set.

Number	Name	Type
ENSG00000253174	RP11-360L9.7	antisense
ENSG00000261560	RP11-166B2.3	sense_intronic
ENSG00000250608	RP11-933H2.4	processed_transcript
ENSG00000280710	RP11-214F16.8	lincRNA
ENSG00000246790	RP11-730K11.1	antisense
ENSG00000269570	CMB9-55A18.1	antisense

Table 2. Significant genes in the COAD data set.

Number	Name	Type
ENSG00000253161	RP11-150O12.1	lincRNA
ENSG00000263235	RP11-461A8.4	sense_overlapping
ENSG00000271133	CTA-293F17.1	antisense
ENSG00000231768	LINC01354	lincRNA
ENSG00000232386	RP11-66B24.2	lincRNA
ENSG00000253414	RP11-150O12.6	sense_intronic
ENSG00000255318	RP11-655M14.13	lincRNA
ENSG00000226699	RP11-122K13.7	antisense
ENSG00000245648	RP11-277P12.20	antisense
ENSG00000255503	RP11-113K21.4	lincRNA
ENSG00000259125	RP11-545N8.3	antisense
ENSG00000259306	RP11-108K3.2	lincRNA
ENSG00000270104	RP11-245P10.6	lincRNA
ENSG00000266835	LINC01540	lincRNA

3 Methods and results

3.1 Numerical experiments of existing models

3.1.1 Cox proportional hazards regression model

Based on the selected significant factors, this article constructs a multi-factor model for each significant factor. Multivariate Cox regression analysis was performed to obtain more significant factors (Fig. 1). The heat map shows the correlation of the expression of more significant factors, which simplifies the regression model.

In the ROC curve, the points where the model training effect is best are listed. The AUC values of READ and COAD are 0.77 and 0.93, respectively. It can be seen that the training effect of the model is generally good, and the model training effect of the COAD data set is very good, reaching a height of 0.93. In order to show the impact of the model on survival analysis more clearly, a cumulative hazard function diagram is drawn (Fig. 2). Based on the *cut value*, it is divided into high-risk and low-risk categories. When it is greater than the *cut value*, it is high-risk; otherwise, it is low-risk.

High-risk patients have higher mortality and lower survival rates than low-risk patients (Fig. 2). The yellow color area and the blue color area in the figure are clearly demarcated. Among them, in the training sample, high-risk patients have a survival rate of 0 after a survival period of about 1200 days. Moreover, the *p* value of both is less than 0.0001, indicating that the effect of the trained model is more significant.

In order to further evaluate the effect of the model, further predictions are made on the model on the test data set. The AUC value, cumulative risk function graph, and ROC curve of two data sets of rectal cancer and colon cancer are obtained (Fig. 3).

Although the two models have obvious yellow-blue curves in the figure, the training effect of the two models on the test data set has decreased to a certain extent, and the *p* value has also decreased. The AUC values are 0.682 and 0.618 respectively, both of which are greater than 0.5. The cumulative risk chart clearly shows that the survival period of high-risk groups is shorter. To a certain extent, it shows that the model has a certain predictive type, but the predictive effect is not very satisfactory. There are many reasons for this phenomenon, which contains less data is an objective reason, the choice of parameters is also very important. Another reason is that the training and nature of the model itself determine the quality of its fit. To this end, this article will continue to train the ensemble learning model on the two types of data.

3.1.2 Random Survival Forest Model

The two data sets were used to establish a random forest model for survival analysis. In order to determine the optimal range of parameters for the random survival forest model, the overall large-scale search for the influence of the model parameters on the results, the size of the tree is preset to 1000, and the size of the different tree models contributes differently to the error rate.

The more trees there are, the smoother the model, the more stable the error rate, and the more accurate the trained model, but the complexity of the model corresponding to the threshold increases. For ensemble

learning, especially random forest, the more likely it is to cause over-fitting. Therefore, in order to balance the gap between the two, select the smallest error rate and the number of trees in the middle of the tree as the number of trees. Based on the above principles, we selected 800 trees as the node size of the READ and COAD data sets for training, and the error rates were 26.49% and 23%, respectively. On the READ data set, the model has the best training effect, the *p* value is small, and there is a clear boundary between the high-risk patient area and the low-risk patient area. In addition, in less than two years, the survival number of high-risk patients has reached zero. On the COAD data set, the distinction between the high-risk patient area and the low-risk patient area is more obvious than that on the READ data set. But the difference is those high-risk patients have a longer survival time, but the *p* value is small, and the training effect of the model is still significant (Fig. 4).

According to the value of AUC (Table 3), it can be found that the model training effect is better in the case of 800 trees. During the three-year life span, the effect of the model reached about 97%.

Based on this model, this paper evaluates the formed model on the testing set. Although the READ data set is very good in training effect, but in the test effect, the effect of *p* value is not ideal. The possible reason is that there are too few training samples, there are only 158, and there are few features after dimensionality reduction. Similarly, although COAD can distinguish between high and low risks, the model training effect is not significant due to the large *p* value. Integrated learning methods such as random forests are particularly suitable for training with large samples and training with more features, and are prone to overfitting. Although the training effect of the model is not particularly significant, it also has a certain degree of predictability (Table 4).

Table 3. The performance of READ and COAD on the training data set.

	Time	300	600	900
AUC Values	READ	1.0000000	0.9273358	0.9761507
	COAD	0.9784992	0.9746421	0.9466558

Table 4. The performance of READ and COAD on the test data set.

	Time	300	600	900
AUC Values	READ	0.6119403	0.7120544	0.7090359
	COAD	0.6272941	0.6464688	0.6562995

Fig. 1. Significant factor expression correlation. (A) Heat map of READ significant factor expression correlation. (B) Heat map of the correlation of COAD significant factor expression.

Fig. 2. The effect of the training model on two types of data sets. (A) ROC curve of READ training model. (B) ROC curve of COAD training model. (C) Survival curve of READ training model. (D) Survival curve of COAD training model.

Fig. 3. The effect of the two types of data sets on the prediction model. (A) ROC curve of READ prediction model. (B) ROC curve of COAD prediction model. (C) Survival curve of READ prediction model. (D) Survival curve of COAD prediction model. (E) READ prediction model cumulative risk diagram. (F) COAD prediction model cumulative risk diagram.

Fig. 4. The prediction of random survival forest and the effect of training model. (A) The effect of READ in training the model. (B) The effect of COAD in training the model. (C) READ test model effect. (D) COAD test model effect.

3.1.3 Optimized random survival forest model

Firstly, optimize the READ data set model. Because the size of the tree in the READ data set is 800, and the amount of data is not too large and the features are small, over-fitting occurs. It is necessary to reduce the complexity of the model while ensuring that the generalization error is as small as possible. For this reason, the size of the tree is optimized so that the size of the tree is still kept to a minimum while ensuring the minimum error rate. Therefore, the model is further refined to keep the size of the tree between 0 and 200 (Fig. 5). When the size of the tree is 105, the error rate is the lowest, which is regarded as *ntree*. Secondly, optimize the COAD data set model. According to the error rate graph, the scale of the tree in COAD was reduced, and finally, 400 was selected as the value of *ntree* of the tree.

Fig. 5. The relationship between the optimized tree model and the error rate. (A) Preliminary refinement of the READ model. (B) The relationship between tree model and error rate after READ refinement model. (C) Preliminary refinement of the COAD model. (D) The relationship between tree model and error rate after READ refinement model.

Train the model again. The effect of the READ training model is still in the best state. When evaluated on the test data set, the model effect is improved compared to the model before optimization, and its AUC value reaches 0.77. When the survival period is 900 days, the prediction effect is increased by 0.01% compared with the model before optimization, and the *p* value is significantly reduced; the optimization effect is relatively improved (Fig. 6). COAD is still as significant as the model before optimization. The low-risk survival rate exceeds 80%, and the high-risk survival rate is less than 25%. There is a clear distinction relationship. And the AUC value reaches more than 0.96 in the period of 2 to 3 years. Compared with the model before optimization, the test model has a certain improvement. When the survival period is three years, the AUC value is 0.647, which is 0.01 compared with the model before optimization, but the *p* value is 0.0637.

3.2 RCCT method integration

The RCCT method is based on the random survival forest model of the integrated learning method and is obtained by integrating the Cox survival model again. It makes up for the shortcomings of the a priori hypothesis of the survival analysis method and the easy overfitting of the random forest method. The value of the risk function of a single sample is shown in equation (1).

$$Obj(t) = \theta H(t) + (1 - \theta)h(t) \quad (1)$$

Among them, t represents a single sample, $H(t)$ is the hazard function value of a single sample of the random survival forest model, and $h(t)$ is the hazard function value of a single sample of the Cox survival model. In addition, the *cutoff* value of the Cox survival analysis method is combined with the *cutpoint* value of the random forest model, as shown in equation (2).

$$cut = \theta cutpoint_H + (1 - \theta) cutoff_h \quad (2)$$

Among them, θ is the coefficient value of the model proportion, and the range is between 0 and 1, which is obtained through the later experimental link.

On the basis of the optimized model in the previous section, in order to obtain a better model effect, the model is improved to find the θ in equations (1) and (2). In the experiment, the interval is 0.01 in the range of 0 to 1, and the general range of the θ value is found through the combination of visual graphics and AUC value judgment, and the initial range is determined to be about 0.9. Then the θ is further refined, with an interval of 0.001 between 0.8 to 0.9 and 0.9 to 1, and then subdivided in this way to determine the specific range of θ . Finally, the experimental results show that when the parameter is 0.8997, the model has a better effect.

The *p* value of the improved method training and the prediction model in READ are numerically equal (Fig. 7). However, in combination with the AUC value in Table 3-6, it can be seen that the AUC value has increased by 0.02. So overall, the effect of the improved model has been improved (Table 5). On the COAD data, the effect of the improved training model is still very good, reaching an AUC value of 0.97. Compared with the previous model, the prediction effect is reduced by 0.003, and the AUC value has increased by 0.03 (Table 6). Compared with the previous model, the effect of the improved model has been further improved. Prove the validity of the model. To a certain extent, the accuracy of the forecast is improved.

Table 5. Performance of READ in test and training data sets after optimization.

	Time	600	900	1200
AUC Values	Training	0.9060897	0.9761507	0.9786336
	Test	0.7273919	0.7365714	0.7970347

Table 6. Performance of COAD in test and training data sets after optimization.

	Time	600	900	1200
AUC Values	Training	0.972554	0.9500573	0.9245709
	Test	0.6579195	0.671244	0.6064147

Fig. 6. The training and testing effects of the optimized model. (A) Training survival curve after READ refines the model. (B) The predicted survival curve after READ refined model. (C) Training survival curve after COAD refines the model. (D) The predicted survival curve after COAD refinement model.

Fig. 7. Survival curve chart of the prediction and training effect after the improved model. (A) Training survival curve after READ improved model. (B) The predicted survival curve after READ improved model. (C) Training survival curve after COAD improved model. (D) The predicted survival curve after the COAD improved mode.

4 Conclusions

This research established a survival analysis prediction model based on ensemble learning methods. Firstly, cleaning for colorectal cancer, colon cancer transcriptome lncRNA and clinical data, dimensionality reduction, and integration, screened significant factors associated with the two data sets. Then each of its survival analysis, survival analysis was used traditional methods, integrated learning random survival forest, the data training model. During the experiment, it is found that the random survival forest method is better than the Cox survival analysis on the training data set, and its AUC value is higher than the Cox method by 0.05. In the performance on the test set, the effect of the random survival forest method is significantly higher than that of the Cox survival analysis, but the highest AUC value of the random survival forest is 0.71. The model is further improved and optimized to increase its AUC value by 0.06. On this basis, the Cox regression model and risk value are integrated again, and the RCCT method is proposed. The experimental results show that the training effect of this model is not much different from the training effect of the random survival forest model. But the effect of this model on the test set is 0.03 higher than the AUC value of the random survival forest model. This shows the effectiveness of the model.

The innovations of this research are mainly reflected in the following aspects. Firstly, based on the two cancer data sets, the significant lncRNA factors that affect survival have been identified, which is conducive to further research on survival-related genes and therapeutic targets of the two cancers in the future. Secondly, based on the random survival forest model, it is integrated with the Cox proportional hazard regression model again, and the risk values of the two are combined to find a better survival prediction method. Improvements or prospects for follow-up work are mainly reflected in the following aspects. First of all, because the data has only a few hundred samples, especially rectal cancer data, the training data has only 75 samples, and the sample size of the data is small, which is not enough to explain the universality of the model. More data may be needed to test the effect of the model. In addition, to explore the mechanism of cancer data sets on a deeper level, based on existing research, we can explore and analyze multiple omics data such as methylation, miRNA, etc., explore the relationship between multiple omics, and find out more. Significantly expressed factors based on omics. Furthermore, since the data set does not provide features such as age, gender, region, etc., the influence of other factors on the two cancers is not ruled out. This requires the collection of other factors to further explore the impact mechanism of the two cancers, to better provide reference recommendations for clinical medicine.

Acknowledgement

This study was supported by 2019 National College

Student Innovation and Entrepreneurship Training Program Project (Approval No. 201914438005).

References

1. Liu Changdong. Research on analysis and prediction of delay in diagnosis and treatment of breast cancer based on statistical learning[D]. Dalian University of Technology, 2019.
2. Li Yongmin. Research on recognition of long non-coding RNA based on machine learning[D]. Nanjing University of Posts and Telecommunications, 2020.
3. Liu Yi. Research on cancer diagnosis method based on machine learning[D]. Hubei University of Technology, 2017.
4. B. Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies[J]. Philosophical transactions of the Royal Society of London, 1825, 115: 513-583.
5. Liu Xin. Research on machine learning and its application in survival analysis[D]. University of Electronic Science and Technology of China, 2018.
6. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests[J]. Ann Appl Stat, 2008, 2(3): 841-860. DOI: 10.1214/08-aos169.
7. Sang Haokai, Guo Shuli, Qu Hong, Zhao Min, Qu Dacheng. Using omics data to build a survival prediction model for four female cancers based on machine learning methods[J]. Science in China: Life Sciences, 2019, 49(06): 738-748.
8. Li Tian. Research on colorectal cancer prediction model based on feature selection and ensemble learning[D]. Southwest University, 2017.
9. Xi Gu, Ziyu Xu, Yiting Liu, Zonghang Liu, Lifeng Zheng. Construction of competing endogenous RNA network and identification of novel molecular biomarkers in colon cancer: A bioinformatic analysis[J]. Medicine, 2021, 100(14).
10. Zhang Qiaosheng. Research on cancer biological pathway analysis method based on omics data[D]. Harbin Institute of Technology, 2019.