Development of a Prognostic Model for Asian Cancer Patients Using Bayesian Networks Based on the SEER Database: A Case Study of Non-Small Cell Lung Cancer (Postprint)

Authors: Yin Bincan, Xin Shichao, Zhang Han, Zhao Yuhong

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: To utilize the SEER database to identify prognostic factors affecting survival in non-small cell lung cancer patients and predict their prognostic survival status, thereby guiding tumor prognostic evaluation.

Methods: Univariate statistical methods and Logistic regression analysis were employed to preliminarily screen prognostic-related factors. The Bayesian network method was utilized to construct a postoperative survival prediction model for patients, and its performance was compared with models built by three other common machine learning classification algorithms.

Results: A total of 5 prognostic variables were ultimately included in the model, including age, tumor size, histological grade, tumor stage, and ratio of involved lymph nodes. The Bayesian network model achieved a prediction accuracy of 72.87% for survival status in non-small cell lung cancer patients.

Limitations: The prognostic factors included in the SEER database are limited, which affects the prediction performance to a certain extent.

Conclusion: Bayesian networks can explore relationships between variables and construct an optimal prognostic model for lung cancer patients, assisting physicians in assessing patient prognosis and treatment efficacy, and is superior to the three models of decision tree, support vector machine, and artificial neural network.

Full Text

Building Asian Tumor-Patients Prognostic Model with Bayesian Network and SEER Database—Case Study of Non-Small Cell Lung Cancer

Yin Bincan[1], Xin Shichao[1], Zhang Han[1], Zhao Yuhong[1,2]
[1](Department of Medical Informatics, China Medical University, Shenyang 110122, China)
[2](Shengjing Hospital of China Medical University, Shenyang 110004, China)

Abstract

[**Objective**] This study aims to identify prognostic factors affecting the survival of non-small cell lung cancer (NSCLC) patients and predict their prognostic status using the SEER database, thereby guiding tumor prognostic assessment. [**Methods**] We initially screened prognostic-related factors using univariate statistical methods and Logistic regression analysis, then constructed a postoperative survival prediction model for patients using the Bayesian network method, and compared its performance with models built by three other common machine learning classification algorithms. [**Results**] A total of five prognostic variables were ultimately included in the model: age, tumor size, histological grade, tumor stage, and lymph node ratio. The Bayesian network model achieved a prediction accuracy of 72.87% for the survival status of NSCLC patients. [**Limitations**] The SEER database includes a limited number of prognostic factors, which may affect the prediction performance to some extent. [**Conclusions**] Bayesian networks can explore relationships between variables and construct optimal prognostic models for lung cancer patients, assisting physicians in evaluating patient prognosis and treatment outcomes. This approach is superior to the three paradigms of decision trees, support vector machines, and artificial neural networks.

Keywords: Bayesian Networks, Non-Small Cell Lung Cancer, Prognosis, Machine Learning

Introduction

Lung cancer is the leading cause of cancer-related mortality, with non-small cell lung cancer (NSCLC) accounting for approximately 83% of all lung cancer cases. The incidence rate is 40.60 per 100,000, and the five-year survival rate is only 22.1%[1]. Given the high incidence and poor prognosis of NSCLC, accurate prognostic assessment is particularly crucial. Currently, clinicians typically evaluate prognosis based on surgical pathological staging, which only considers three aspects: primary tumor site, regional lymph node involvement, and distant metastasis, while ignoring other prognostic factors, resulting in poor predictive performance[2]. Existing prognostic studies are mostly conducted in single or a few medical institutions, with significant missing follow-up data, small sam-

ple sizes, and low credibility. There is an urgent clinical need for a prognostic prediction and evaluation system for NSCLC patients based on larger datasets with high credibility and good predictive performance.

The National Cancer Institute (NCI) established the Surveillance, Epidemiology, and End Results (SEER) database in 1973, which is internationally recognized as an authoritative source of cancer patient follow-up data and provides reliable data support for clinical research. Some scholars have used this database to establish survival prediction models for diseases such as rhabdomyosarcoma using simple statistical methods. This study will utilize the SEER database to extract NSCLC cases among Asian populations and employ machine learning methods that better reflect the correlations between prognostic variables and offer better applicability to construct a prognostic model and prediction evaluation system for Asian NSCLC patients, providing decision support for clinical treatment and prognostic assessment.

Both domestic and international research on disease prediction models has established a certain foundation. Muers et al.[3] collected NSCLC patient data from six medical institutions to establish a prognostic risk model and compared the model' s predicted survival with clinicians' judgments. Yang et al.[4] constructed five-year and ten-year survival prediction models for rhabdomyosarcoma patients based on the SEER database to guide treatment selection. Park et al.[5] used clinical trial data to predict survival in patients with advanced biliary tract adenocarcinoma receiving palliative chemotherapy. All these models first screened prognostic factors and then used COX regression methods from statistics to construct the models, which is also a common approach for building medical prediction models. However, COX regression analysis makes it difficult to visualize relationships between prognostic variables. To improve model applicability, machine learning methods have gradually gained researchers' favor. For instance, a nomogram based on seven indicators can determine the likelihood of postoperative recurrence. One researcher used support vector machine methods in 2012 to predict five-year survival in breast cancer patients[7] and subsequently built an online prognostic system.

Since the early 21st century, an increasing number of domestic researchers have begun to evaluate the occurrence, development, and prognosis of tumors and other diseases from a machine learning perspective. Liu Yaqin[8] compared prognostic prediction models using Logistic regression, artificial neural networks, and decision trees based on the SEER database, representing an important breakthrough in tumor prognosis research in this field in China. Taiwanese scholar Chen et al.[9] used artificial neural networks to investigate clinical and gene expression data from NSCLC patients in four medical institutions and established a survival risk model. Mu Dongmei et al.[10] extracted electronic medical record information to construct a risk factor prediction model for pregnancy-induced hypertension syndrome and found the decision tree model to be optimal. However, variable selection in these studies relied solely on existing experience, lacked communication with clinicians, and did not achieve interdisciplinary col-

laboration.

Through literature review, we found that prognostic studies on lung cancer, which has high incidence and mortality rates, are scarce. Therefore, this study, based on the SEER database, identifies patient prognostic factors and refines them with input from oncologists. By employing machine learning methods that better reflect correlations between prognostic variables and offer greater applicability, we aim to improve prediction accuracy and construct a postoperative survival model for Asian NSCLC patients to better serve clinical prognostic evaluation.

3. Tumor Prognosis Model Construction Scheme

Tumor prognosis includes risk assessment, recurrence, metastasis, and survival evaluation[11]. Using five years post-surgery as the time benchmark for NSCLC patients, we predict patient survival status (survival vs. death). The specific research flow is shown in Figure 1 [Figure 1: see original paper].

Figure 1. Research Flow for Building Asian NSCLC Patient Prognostic Model Based on SEER

The specific steps are as follows:

(1) **Data Download:** In SEER*Stat software, we accessed the Incidence-SEER18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2014 version, with follow-up data ending in late 2012. NSCLC patient data were downloaded according to ICD-O-3 morphology codes for malignant tumors.

(2) **Variable Selection Basis:** Referencing prognostic factors related to patient survival mentioned in the American Joint Committee on Cancer (AJCC), National Comprehensive Cancer Network (NCCN) clinical guidelines, and the U.S. Collaborative Stage Manual Online Help (CS) system[12-13], we extracted all fields containing these variables from SEER*Stat. Using patient information recorded at initial diagnosis, we organized the data into Excel spreadsheets.

(3) **Feature Variable Screening:** To determine whether each variable independently affected patient survival, we first performed univariate analysis (independent samples t-test or chi-square test) on training samples using SPSS 22.0. Variables identified through univariate analysis were then included in Logistic regression analysis to screen for highly relevant prognostic factors in NSCLC ($P < 0.05$ was considered statistically significant). Variables were adjusted based on clinical physicians' recommendations for inclusion in the final model.

(4) **Tumor Prognosis Model Construction:** We employed supervised learning methods in machine learning to construct the tumor prognosis prediction model[10]. Using R Studio software, we established a Bayesian sur-

vival prediction model, completed structural adjustments to the Bayesian network, and built an effective prognostic model.

(5) **Model Evaluation:** We used the data mining software WEKA to compare the prediction accuracy, precision, and area under the ROC curve of the Bayesian network model with three other common classification models.

4.1. Construction of Tumor Prognosis Model

(1) **Study Subjects:** Asian patients diagnosed with NSCLC from 2004 onward were selected as the final study subjects, including patients who died directly from NSCLC within five years and those who survived for the full five-year follow-up period, totaling 683 cases.

(2) **Study Variables:** Seventeen prognostic variables were extracted from SEER: gender, nationality, marital status, primary site, histology type, histological grade, laterality, degree of adjacent organ infiltration, degree of regional lymph node involvement, degree of distant metastasis, tumor stage, surgery type, radiation therapy receipt, age at diagnosis, tumor size, number of positive lymph nodes, and number of examined lymph nodes. The last four indicators were continuous variables, while the rest were categorical variables, as shown in Table 1 .

Table 1. Prognostic Indicators Information for Non-Small Cell Lung Cancer Patients

SEER Field Name	Categories/Value Range
Race recode (Asian)	
Marital status at diagnosis	
Primary Site - labeled	
ICD-O-3 Hist/behav, malignant	
Histological grade (Grade)	
Regional lymph nodes (Laterality)	
CS extension	
CS lymph nodes	
Distant metastasis degree (CS mets at dx)	
Derived AJCC Stage Group	
RX Summ–Surg Prim Site	
Radiation	
Age at diagnosis	
CS tumor size	
Regional nodes positive	
Regional nodes examined	

(3) **Outcome Variables:** Five-year survival is a critical indicator for evaluating prognostic outcomes. We used five-year postoperative survival status

of NSCLC patients as the dependent variable. Survival time was measured in months and converted to a categorical variable: patients with survival time of 60 months or more were considered "survived" (coded as 1), while others were considered "deceased" (coded as 0).

(4) **Feature Variable Selection:** To reduce the number of prognostic variables and improve model prediction accuracy, we selected highly relevant prognostic factors from the study variables. Variables initially included after univariate analysis ($P < 0.05$) were: age at diagnosis, tumor size, histological grade, tumor stage, degree of adjacent organ infiltration, degree of regional lymph node involvement, number of positive lymph nodes, marital status, nationality, degree of distant metastasis, surgery type, and radiation therapy receipt. Logistic regression analysis based on the univariate analysis further identified the following prognostic variables ($P < 0.05$): age at diagnosis, tumor size, histological grade, tumor stage, number of examined lymph nodes, and number of positive lymph nodes. The screening results are shown in Table 2 .

Table 2. Variable Screening Results from Logistic Regression Analysis

Variable	Exp(B)	95% CI for Exp(B)	
		Lower	Upper
Age at diagnosis	-0.066		
Histological grade			
Regional nodes examined			
Regional nodes positive			

The lymph node ratio (LNR), defined as the ratio of positive lymph nodes to examined lymph nodes, was used as a prognostic variable in place of the two separate lymph node counts based on clinical recommendations. The final variables entered into the model were: age at diagnosis, tumor size, histological grade, tumor stage, and lymph node ratio.

(5) **Data Preprocessing:** We deleted records with severe data missingness, recording errors, or deaths from causes other than lung cancer. The Interval method was used to discretize numerical data. This discretization method aims to divide the interval [X_{min}, X_{max}] into equally sized subintervals D and provide discretization indices based on the subinterval index, where observation index i and discretization level j satisfy the following conditions[14]:

In R Studio, we used the bnlearn package to implement these data preprocessing steps. The data were then split into training ($N_1 = 495$) and test sets ($N_2 = 188$) at approximately a 70:30 ratio[15]. The training set was used for network learning and adjustment to construct the prognostic model, while the test set was used to evaluate model performance.

(6) **Prognostic Model Construction and Prediction Results:** A Bayesian Network (BN) describes the dependency relationships between child and parent nodes through nodes representing variables and connections representing relationships between variables[16]. Given random variables X = {X_1, X_2, ···, X }, their joint probability distribution is:

$$P(X) = \prod_{i=1}^{n} P(X_i|\mathrm{Pa}(X_i))$$

where Pa(X_i) is the subset of parent nodes of X_i, and in the network graph, X_i is independent of variables that are not its direct ancestors. We used the Tabu Search (TS) method for initial Bayesian network learning. Proposed by American Academy of Engineering member Fred Glover in 1986[17], TS is a heuristic algorithm based on neighborhood search and iteration for solving optimization problems. Its essence is to prohibit repeating previous work and escape local optima by randomly moving in the region and generating new solutions, then evaluating each neighboring solution and selecting the path that most improves the objective function. If no solution can improve the final result, the solution with the least impact on the objective function is selected, using human memory imitation to find the optimal result[18]. The steps are as follows:

Determine the neighborhood N(x), select an initial feasible solution X_0 from it, set the current best solution X_{best} = X_0, and let T = N(X_{best});

Combine sequentially according to the above steps to obtain a new solution nextX_n $N^+(\infty)$, and output the calculation results;

Compare all decision results and output the globally optimal decision solution.

Makond et al.[19] constructed a Bayesian prognostic model not entirely based on data learning but by listening to physicians' opinions to build a patient prognostic survival model, which is essentially an experience-based modeling approach. Our study overcomes the limitation of relying solely on experience-based modeling by combining the TS network learning method with physician input to establish a patient prognostic model. Network model refinement and optimization were implemented in R Studio, with the final network model shown in Figure 2 [Figure 2: see original paper].

Using the caret package in R Studio, we output prediction tables composed of prediction samples and instances along with model evaluation metrics. Among the 188 test set samples, 137 were predicted correctly, achieving a prediction accuracy of 72.87%.

Figure 2. Bayesian Network Model for Prognostic Survival of Asian Non-Small Cell Lung Cancer Patients

4.2. Comparative Experiment

We also used decision tree, support vector machine, and artificial neural network methods to build prognostic models and compared their prediction results with our model. In WEKA, we selected J48, SMO, and Multilayer Perceptron corresponding to the three methods, using default parameters. The prediction accuracy and model performance comparisons of the four machine learning algorithms are shown in Tables 4 and 5 .

Table 4. Prediction Accuracy Comparison Between BNNSCLC Model and Three Other Classification Algorithms

Classification Algorithm	Prediction Accuracy
Bayesian Network	72.87%
Support Vector Machine	67.02%
Artificial Neural Network	68.62%
Decision Tree	64.89%

Table 5. Performance Comparison of Models Built by Different Algorithms

Algorithm	Prediction Accuracy	Precision	AUC-ROC
Bayesian Network	72.87%	71.0%	0.68
Support Vector Machine	67.02%	66.3%	0.66
Artificial Neural Network	68.62%	68.2%	0.63
Decision Tree	64.89%	63.7%	-

4.3. Experimental Analysis

This study found that the NSCLC prognostic model constructed by Bayesian network is optimal. As shown in Table 4, although decision tree, support vector machine, and artificial neural network achieved higher prediction accuracy on the training set than the Bayesian network, their prediction accuracy on the test set decreased significantly compared to the training set. They failed to adapt well to new data and are not suitable for practical application, indicating poorer model fitting than the Bayesian network model. Furthermore, interpreting Table 5 reveals that the Bayesian network model outperformed the other three machine learning models in terms of prediction accuracy, precision, and area under the ROC curve.

The selection of network learning methods is fundamental to building Bayesian classifiers. This study used the TS method for initial network model construction, which is an optimization of hill climbing. When it is known that certain network variables do not create network loops, TS uses move search instead of

random generation, employing three operations—adding, deleting, and reversing edges—to generate neighborhoods[20] and search for global optimal solutions to adjust network structure and complete Bayesian network self-learning. On this basis, we combined clinical physicians'experience to modify the network diagram by linking highly relevant prognostic factors, representing a typical combination of theoretical methods and practical application.

Network diagram adjustment is the most critical process in constructing this survival prediction model. As shown in Figure 2, arrow directions indicate relationships between nodes. For example, size pointing to stage indicates that the former directly influences the latter. All selected prognostic variables point to the final variable—survival status—among which age at diagnosis, tumor stage, and lymph node ratio directly affect patient survival.

By constructing different network diagrams to find the optimal classification model, we can determine relationships between prognostic factors and their impact on survival status. Clinicians can use this to evaluate postoperative prognosis in cancer patients and control relevant factors. However, since the SEER database used in this study does not include all tumor prognostic factors[21], the number of indicators available for modeling is limited, which may introduce certain limitations to the prediction model.

This study constructed a prognostic survival model for NSCLC patients with postoperative survival status as the target, achieving a prediction accuracy of 72.87%. By building a Bayesian network to explore relationships between prognostic variables and their impact on patient survival, and by incorporating clinical expert recommendations based on internal network structure adjustments, we better interpreted the relationships between nodes in the model. For the first time, we used the SEER database to construct a survival prediction model focusing on Asian cancer patients, which can assist in evaluating patient prognosis five years post-surgery and shows promising application prospects. Future research could consider incorporating external validation from other patient sources to improve the model' s adaptability and better serve clinical treatment and prognostic evaluation.

References

[1] National Cancer Institute. SEER Cancer Statistics Review (CSR) 1975-2013 [R/OL]. [2016-09-20]. http://seer.cancer.gov/csr/1975_{2013}/sections.html.

[2] Ettinger D S, Wood D E, Akerley W, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 4.2016 [J]. Journal of the National Comprehensive Cancer Network: JNCCN, 2016, 14(3): 255-264.

[3] Muers M F, Shevlin P, Brown J. Prognosis in Lung Cancer: Physicians' Opinions Compared with Outcome and a Predictive Model[J]. Thorax, 1996, 51(9): 894-902.

[4] Yang L, Takimoto T, Fujimoto J. Prognostic Model for Predicting Overall

Survival in Children and Adolescents with Rhabdomyosarcoma[J]. BMC Cancer, 2014, 14: 654. DOI: 10.1186/1471-2407-14-654.

[5] Park I, Lee J L, Ryu M H, et al. Prognostic Factors and Predictive Model in Patients with Advanced Biliary Tract Adenocarcinoma Receiving First-line Palliative Chemotherapy[J]. Cancer, 2009, 115(18): 4148-4155.

[6] Kim W, Kim K S, Park R W. Nomogram of Naive Bayesian Model for Recurrence Prediction of Breast Cancer[J]. Healthcare Informatics Research, 2016, 22(2): 89-94.

[7] Kim W, Kim K S, Lee J E, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine[J]. Journal of Breast Cancer, 2012, 15(2): 230-238.

[8] Liu Yaqin. Study on the Prognosis Model for Breast Cancer[D]. Shanghai: Shanghai Jiaotong University, 2008.

[9] Chen Y C, Ke W C, Chiu H W. Risk Classification of Cancer Survival Using ANN with Gene Expression Data from Multiple Laboratories[J]. Computers in Biology and Medicine, 2014, 48: 1-7.

[10] Mu Dongmei, Ren Ke. Discovering Knowledge from Electronic Medical Records with Three Data Mining Algorithms[J]. New Technology of Library and Information Service, 2016(6): 102-109.

[11] Shin H, Nam Y. A Coupling Approach of a Predictor and a Descriptor for Breast Cancer Prognosis[J]. BMC Medical Genomics, 2014, 7(S1): S4.

[12] American Joint Committee on Cancer. AJCC Cancer Staging Manual[M]. The 7th Edition. New York: Springer Verlag, 2010: 253-270.

[13] National Comprehensive Cancer Network: NCCN Clinical Practice Guidelines in Oncology: Non-Small Cell Lung Cancer, Version 2.2016[R/OL]. [2016-09-20]. http://www.nccn.org/patients.

[14] Hartemink A J. Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks[D]. Massachusetts Institute of Technology, 2001: 86-87.

[15] Kumar Y, Sahoo G. Prediction of Different Types of Liver Diseases Using Rule Based Classification Model[J]. Technology & Health Care Official Journal of the European Society for Engineering & Medicine, 2013, 21(5): 417-432.

[16] Oh J H, Craft J, Al L R, et al. A Bayesian Network Approach for Modeling Local Failure in Lung Cancer[J]. Physics in Medicine & Biology, 2011, 56(6): 1635-1651.

[17] Zhang Xuelei. The Application of Bayesian Network Based on Tabu Search Algorithm in Diseases Prediction and Diagnosis[D]. Taiyuan: Shanxi Medical University, 2015.

[18] Lim W L, Wibowo A, Desa M I, et al. A Biogeography-Based Optimization Algorithm Hybridized with Tabu Search for the Quadratic Assignment Problem[J]. Computational Intelligence & Neuroscience, 2016. DOI: 10.1155/2016/5803893.

[19] Makond B, Wang K J, Wang K M. Probabilistic Modeling of Short Survivability in Patients with Brain Metastasis from Lung Cancer[J]. Computer Methods & Programs in Biomedicine, 2015, 119(3): 142-162.

[20] Wei Zhen, Zhang Xuelei, Rao Huaxiang, et al. Using the Tabu-search-algorithm-based Bayesian Network to Analyze the Risk Factors of Coronary Heart Diseases[J]. Chinese Journal of Epidemiology, 2016, 37(6): 895-899.

[21] Yang Qiao, Zhang Junping. Clinical Applications of the Tumor Registry Database[J]. The Journal of Evidence-Based Medicine, 2013, 13(4): 250-251, 256.

Author Contributions Statement

Yin Bincan: Designed the research protocol, performed data analysis, constructed the model, and wrote the manuscript; Xin Shichao: Performed data preprocessing and modeling experiments; Zhang Han: Revised the manuscript; Zhao Yuhong: Proposed the research idea and revised the final version of the manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, E-mail: yinbincan0803@163.com.
[1] Yin Bincan. NSCLC.csv. Raw data for Asian non-small cell lung cancer patient prognostic model research.
[2] Yin Bincan. data.csv. Asian non-small cell lung cancer patient modeling data.

Received: October 31, 2016
Revised: December 5, 2016

Note: Figure translations are in progress. See original paper for figures.