Construction 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 identify prognostic factors affecting survival in non-small cell lung cancer patients and predict their prognostic survival status using the SEER database, 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, with its performance compared against models built by three other common machine learning classification algorithms.

【Results】 A total of 5 prognostic variables were finally included in the model, comprising 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 limited prognostic factors included in the SEER database affect the prediction effectiveness to a certain extent.

【Conclusion】 Bayesian networks can explore relationships among variables and construct an optimal prognostic model for lung cancer patients, assisting physicians in evaluating patient prognosis and treatment efficacy, and outperforming the three models of decision tree, support vector machine, and artificial neural network.

Full Text

Building a Prognostic Model for Asian Cancer Patients Using Bayesian Networks and the SEER Database: A 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 leverages the SEER database to identify prognostic factors influencing survival in non-small cell lung cancer (NSCLC) patients and to predict their prognostic status, thereby guiding tumor prognosis assessment. **Methods:** We employed univariate statistical methods and logistic regression analysis to initially screen for prognosis-related factors, then constructed a post-operative survival prediction model using Bayesian network methodology. The performance of this model was compared against three other common machine learning classification algorithms. **Results:** Five prognostic variables were ultimately incorporated into 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 NSCLC patient survival status. **Limitations:** The SEER database includes a limited number of prognostic factors, which may affect prediction effectiveness 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 efficacy. This approach outperforms decision trees, support vector machines, and artificial neural networks.

Keywords: Bayesian Networks; Non-Small Cell Lung Cancer; Prognosis; Machine Learning
Classification Numbers: R730.7; G35

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 of NSCLC is 40.60 per 100,000, and the five-year survival rate is only 22.1% [1]. Given its high incidence and poor prognosis, accurate prognostic assessment for NSCLC is particularly critical. Currently, clinicians typically rely on surgical pathological staging to judge prognosis. However, this staging system only considers three aspects—primary tumor site, regional lymph node involvement, and distant metastasis—while ignoring other prognostic factors, resulting in suboptimal predictive performance [2]. Most existing prognostic studies are limited to single or a few medical institutions, with substantial

missing follow-up data, small sample 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 higher credibility and better 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 utilized this database to develop survival prediction models for diseases such as rhabdomyosarcoma using simple statistical methods. This study will extract Asian NSCLC cases from the SEER database and employ machine learning methods that better reflect correlations between prognostic variables and offer greater applicability to construct a prognostic model and prediction evaluation system for Asian NSCLC patients, thereby providing decision support for clinical treatment and prognosis assessment.

Research on disease prediction models has already established a solid foundation both domestically and internationally. Muers et al. [3] collected NSCLC patient data from six medical institutions to develop a prognostic risk model and compared the model's predicted survival with clinicians' judgments. Yang et al. [4] built 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 advanced biliary tract adenocarcinoma patients receiving palliative chemotherapy. All these models employed COX regression, a common approach in medical prediction modeling. However, COX regression analysis makes it difficult to visualize relationships between prognostic variables. To improve model applicability, machine learning methods have gained increasing recognition among researchers. For instance, one study developed a nomogram using seven indicators to predict postoperative recurrence probability [6], while the same research team previously used support vector machines 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 evaluating 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 domestic tumor prognosis research. Taiwanese scholar Chen et al. [9] used artificial neural networks to investigate clinical and gene expression data from NSCLC patients across four medical institutions, establishing 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, finding decision trees to be optimal. However, variable selection in these studies relied primarily on existing experience without interdisciplinary collaboration with clinicians. Literature review reveals that prognostic

research on lung cancer, which has high incidence and mortality rates, remains scarce. Therefore, this study identifies prognostic factors based on the SEER database, refines them through consultation with oncologists, and employs machine learning methods that better reflect relationships between prognostic variables and offer superior applicability to construct a postoperative survival model for Asian NSCLC patients with the goal of improving prediction accuracy and better serving clinical prognosis evaluation.

3. Tumor Prognosis Model Construction Scheme

Tumor prognosis encompasses risk assessment, recurrence, metastasis, and survival evaluation [11]. Using five-year postoperative survival as the temporal benchmark, this study predicts patient survival status (survival vs. death). The specific research process is illustrated in Figure 1 [Figure 1: see original paper].

Figure 1. Research Process for Building a Prognostic Model for Asian NSCLC Patients 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 ending in late 2012) and downloaded NSCLC patient data according to ICD-O-3 morphology codes for malignant tumors.

2. **Variable Selection Rationale:** 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 second edition of the Collaborative Stage Manual (CS) [12-13], we extracted all fields containing these variables from SEER*Stat. Using patient information recorded at initial diagnosis, we compiled 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 the training sample 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, with $P < 0.05$ considered statistically significant. Variables were adjusted for final model inclusion based on clinical physician recommendations.

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 survival prediction model and completed structural adjustments to build 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. Tumor Prognosis Model Construction

(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 the full five-year follow-up period, totaling 683 cases.

(2) Study Variables

Seventeen prognostic research variables were extracted from SEER: gender, nationality, marital status, disease site, pathological type, histological grade, tumor 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 Indicator Information for Non-Small Cell Lung Cancer Patients

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

(3) Outcome Variable

Five-year survival status is a crucial indicator for evaluating prognosis effectiveness. We used five-year postoperative survival status as the dependent variable. Survival time (in months) was 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 performed selection of highly relevant prognostic factors. 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 results 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 Selection 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 incorporated as a prognostic variable based on clinical recommendations, replacing the two separate variables of positive lymph node count and examined lymph node count. The final variables entering 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, and patients who died 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 the observation index i and discretization level j satisfy the following conditions [14]:

[Mathematical conditions for discretization]

These data preprocessing steps were implemented in R Studio using the bnlearn package. The data were then split into training ($N_1 = 495$) and test ($N_2 = 188$) sets 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 employed the Tabu Search (TS) method for preliminary Bayesian network learning. Proposed by American Academy of Engineering member Fred Glover in 1986 [17], TS is a heuristic algorithm that solves optimization problems based on neighborhood search and iteration. The method essentially prohibits repeating previous work to 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 minimal impact on the objective function is selected, using human memory imitation to find the optimal result [18]. The steps are as follows:

1. Determine the region N(x), select an initial feasible solution X_0 from it, set the current optimal solution X_{best} = X_0, and T = N(X_{best});
2. Combine sequentially according to the above steps to obtain new solutions nextX_n $N^+(\infty)$, and output the calculation results;
3. Compare all decision results and output the global optimal solution.

Makond et al. [19] constructed a Bayesian prognostic model not entirely based on data learning but by incorporating physician opinions to build a patient prognostic survival model, representing an experience-based modeling approach. This study overcomes the limitation of relying solely on experience-based modeling by combining the TS network learning method with physician opinions to construct the 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].

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

In R Studio, the caret package was used to output prediction tables and model evaluation metrics. Among the 188 test set samples, 137 were predicted correctly, achieving a prediction accuracy of 72.87%.

4.2. Comparative Experiment

We additionally constructed prognostic models using decision tree, support vector machine, and artificial neural network methods for comparison with our proposed model. In WEKA, we selected J48, SMO, and Multilayer Perceptron algorithms with default parameters to build the prognostic models. The prediction accuracy and model performance comparisons across 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	
Support Vector Machine	
Artificial Neural Network	

Table 5. Performance Comparison of Models Built with Different Algorithms

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

4.3. Experimental Analysis

This study found that the Bayesian network constructed the optimal NSCLC prognostic model. As shown in Table 4, although decision trees, support vector machines, and artificial neural networks 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, indicating poor adaptation to new data and limited practical applicability. Consequently, these models exhibited inferior fitting compared to the Bayesian network model. Furthermore, Table 5 demonstrates that the Bayesian network model surpassed the

other three machine learning approaches in prediction accuracy, precision, and area under the ROC curve.

The selection of network learning methods is fundamental to building Bayesian classifiers. This study employed the TS method for preliminary network model construction, representing an optimization of hill climbing. When it is known that certain network variables do not create network loops, TS replaces random generation with mobile search, using 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 foundation, we incorporated clinical physician experience to modify the network graph by connecting highly relevant prognostic factors, representing a typical integration of theoretical methods and practical application.

Network graph adjustment constitutes 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 graphs to identify the optimal classification model, we can determine relationships among prognostic factors and their impact on survival status, enabling clinical evaluation of postoperative prognosis and control of relevant factors. However, since the SEER database used in this study does not include all tumor prognostic factors [21], the number of selected indicators is limited, which may constrain the prediction model.

This study constructed a postoperative survival prognostic model for NSCLC patients with 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 into internal network structure adjustments, we better interpreted the relationships between nodes in the model. For the first time, we applied the SEER database to construct a survival prediction model focusing on Asian cancer patients, which can assist in evaluating postoperative five-year prognosis and shows promising application potential. Future research should consider incorporating external validation from other patient sources to enhance the model' s adaptability and better serve clinical treatment and prognosis 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. (in Chinese)

[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. (in Chinese)

[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]. 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. (in Chinese)

[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. (in Chinese)

[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. (in Chinese)

Author Contributions

Yin Bincan: Designed the research protocol, performed data analysis, constructed the model, and wrote the manuscript.
Xin Shichao: Conducted 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 conflicts of interest.

Supporting Data

Supporting data are self-archived by the authors and available upon request at: yinbincan0803@163.com.
[1] Yin Bincan. NSCLC.csv. Raw data for the prognostic model study of Asian non-small cell lung cancer patients.
[2] Yin Bincan. data.csv. Modeling data for Asian non-small cell lung cancer patients.

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

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