Study on Application of Bayesian Networks and SEER Database in the Predicting of Non-Small Cell Lung Cancer

Wei Chen
Department of Computer Science
School of Aerospace Science (University of North Dakota)
Grand Forks, ND, USA, 58202
wei.chen.2@und.edu

Chinedu Nwachukwu
Department of Computer Science
School of Aerospace Science (University of North Dakota)
Grand Forks, ND, USA, 58202
chinedu.nwachukwu@und.edu

Abstract— **Despite various efforts to develop new predictive models for early detection of tumor local failure in locally advanced Non-Small Cell Lung Cancer (NSCLC), many patients still suffer from a high local failure rate after radiotherapy. Based on SEER database, we proposed a graphical Bayesian network framework to identify the influencing factors of Asian patient's survival status and predicts their prognostic situation, and further to improve the tumor-prognostic assessment for the patients who were diagnosed with NSCLC. We used One-Way ANOVA analysis statistical method and logistic regression to identify the prognostic variables, and employed the Bayesian Network algorithm to construct the prognostic survival model for Asian NSCLC patients, as well to compare the performance of our model with three other algorithms. Our experimental results demonstrate that the identified prognostic variables include age, tumor size, grade, tumor stage, as well as the lymph nodes ratio, and the proposed method can be used as an efficient method to predict NSCLC patient's prognostic survival status and to interpret relationships among the different variables. Though the SEER database had limited number of prognostic factors, which may influence the prediction accuracy, only achieving 72.87% at present, the Bayesian Network could help us build optimal prognosis model for cancer patients to improve their survival rates. The proposed model is better than the Decision Tree, Support Vector Machine and Artificial Neural Network models.**

Keywords— Bayesian Networks (BN), Non-Small Cell Lung Cancer (NSCL), Prognosis, SEER Database component)

I. INTRODUCTION

For accurate medical diagnosis and prediction, it is often necessary to model a patient's condition over time. Because there is great uncertainty in clinical medicine, a system for diagnostic or prognostic evaluation must be able to represent and reason with uncertainty [1]. Bayesian Network (BN) models are a class of machine learning models with some unique characteristics that make them suitable for medical applications [2]. First of all, they are currently the most powerful and popular method for representing and reasoning with probabilistic uncertainties inherent to the medical domain. Second, they can be described in the form of a graph making them relatively easily interpretable for the medical community and providing an effective means to reason about new links and graphs. Third, BN models take into account the dependence relations between the features, revealing the cause-effect relationships included in medical data [3].

A Bayesian network has been used as a useful tool to create individualized predictive models due to its several attractive characteristics, which have led to various studies successfully in the field of oncology. Recently, Jayasurya et al. proposed a Bayesian network model for survival prediction in lung cancer patients [4]. They also showed that the Bayesian network can be efficiently used when handling missing data compared with other machine learning techniques. Velikova et al (2009) designed a multi-view mammographic analysis system using a Bayesian network framework to detect breast cancer and demonstrated the potential of the system for selecting the most suspicious cases [5]. Chen et al (2006) proposed an effective Bayesian structure learning method based on the mutual information and K2 algorithm to reconstruct reliable gene networks [6]. Van Gerven et al (2008) demonstrated the development of a prognostic model for carcinoid patients using dynamic Bayesian networks [7]. Smith et al (2009) developed a prognostic model for prostate cancer with intensity modulated radiation therapy (IMRT) plans and calculated a quality-adjusted life expectancy for each plan using Bayesian networks [8].

Lung cancer is a leading cause of cancer death worldwide [9]. Of all lung cancer cases, non-small cell lung cancer (NSCLC) accounts approximately for 83%. The incidence of NSCLC was 40.60/100,000, and 5-year survival rate was only 22.1% [10]. For treating patients with advanced and inoperable stage, a combination of chemotherapy and radiotherapy is mainly used instead of surgical resection [9]. However, patients with locally advanced NSCLC following radiotherapy suffer from a high local failure rate [11]. Despite many efforts to improve treatment outcomes, a low two-year local control rate as low as 27% in these patients requires innovative diagnostic and prognostic models to improve early detection of tumor local failure [12].

The database of Surveillance, Epidemiology and End Results (SEER) was set up by National Cancer Institute (NCI) in 1973. It is one of the world's recognized data sources following up for cancer patients. Providing reliable data support for clinical research. Some scholars had set up rhabdomyosarcoma and other disease survival model by simple statistic method according the SEER database. The present study will use the SEER database to extract the NSCLC cases of Asian; adopt relationship and applicability of better machine learning methods to build Asian people NSCLC prognosis model and prediction assessment system, and provide decision support for the treatment and prognosis of clinicians.

II. BAYESIAN NETWORKS THEORY

BN (Bayesian Networks) defines a unique joint probability distribution over variables. And the joint probability, marginal probability or posterior probability of variables is calculated by probabilistic reasoning. Marginal probability table is tabular form of its distributions that are assigned to root and non-root nodes. BN is a visual graphic method which can express connection between different nodes [5-6].

Bayesian conditional probability definition formula:

$$P(A \mid B) = P(B \mid A)P(A)/P(B) \quad \cdots\cdots\cdots\cdots (1)$$

Where P (B) is prior probability, P(A | B) is posterior probability.

Total probability formula is expressed as:

$$P(B) = \sum_{i=1}^{n} P(B \mid A = a_i)P(A = a_i) \quad \cdots\cdots\cdots\cdots (2)$$

Where n represents the number of state of A. Bi-direction reasoning can be realized by using BN.

III. CONSTRUCTION PLAN OF TUMOR PROGNOSTIC MODEL

The prognosis of tumor includes risk assessment, recurrence, metastasis and survival status evaluation [13]. The survival situation of the patients with NSCLC for five years postoperative time baseline is predicted as "survival" and "death". The process is shown in Figure 1 and the description of the dataset is followed by our proposed approach.

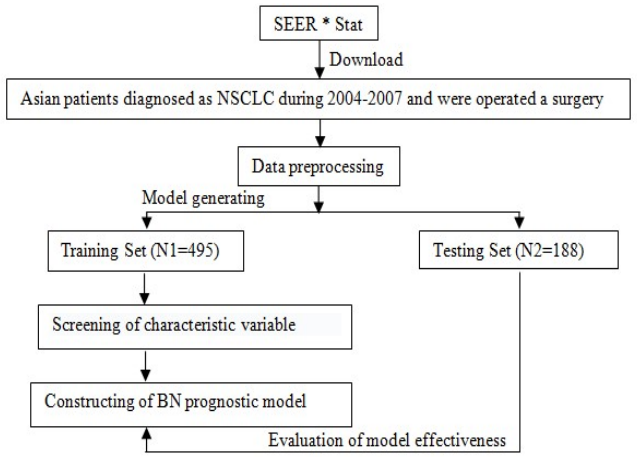

Figure 1 Research process of Asian NSCLC patient's prognostic model constructing based on SEER

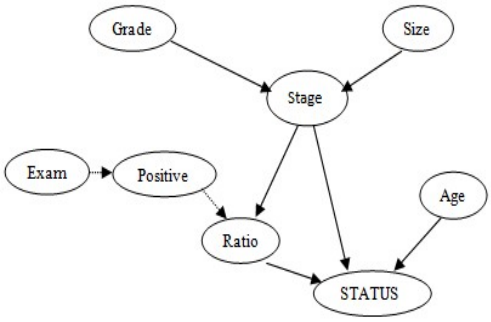

Figure 2 Prognosis survival bayesian network models of Asian non-small cell lung cancer patients

A. Data download

The data of "Incidence-SEER18 Regs Research Data+Hurricane Katrina Impacted Louisiana Cases, Nov2014"was invoked by the STEER*Stat software. The release date of this version was at the end of 2012; NSCLC patient's data was downloaded based on the morphological coding of ICD-O-3 malignant tumor.

B. Basis for the selection of variable

According what mentioned in the American Joint Committee on Cancer (AJCC), the National Comprehensive Cancer Network (NCCN) [14-15], and Collaborative Stage Manual Online Help (CS), related to the patient's prognostic factors, all fields containing the above variables were extracted

from SEER*Stat and registered at the time of the first diagnosis. The information was organized to lead in tables.

C. Screening of characteristic variable

To determine whether the variables are independent in the patient's survival, SPSS22.0 software was first applied to analyze the training samples with One-Way ANOVA (t test), and then Logistic regression analysis was used to analyze the variables obtained by single factor analysis. Screening NSCLC high correlation prognostic factors, P<0.05 was statistically significant. The adjustment variables are incorporated into the final model in combination with the recommendations of clinicians.

D. Construction of tumor prognostic model

Supervised learning method in machine learning was selected to the tumor prognosis prediction model. R Studio software was applied to establish bayesian survival prediction model and to complete the structure adjustment of BN, as well to construct an effective prognostic model.

E. Evaluation of model effectiveness

The data mining software WEKA was selected to compare the predicting accuracy, accuracy and area under the curve of ROC among models of BN and other three common classified.

IV. PROCESSING

A. Construction of tumor prognostic model

1) Research object: Asian patients diagnosed with NSCLC since 2004 were selected as the final study subjects, which included the patient's death of NSCLC within 5 years, and patients who had been followed up for 5 years and still survived. The total number was 683.

2) Research variables: Seventeen prognostic variables were extracted from SEER, as can be seen in Table 1, the last four are continuous variables, and the others are categorical variables.

3) Outcome variable: The five-year survival of NSCLC patients is an important index to evaluate the prognostic effect, and was treated as the dependent variable. One month is made as one survival period unit, and the conversion of the classification variables, Patients who lived for 60 months or more were considered "living" (marked 1), otherwise considered "death" (marked 0).

4) Feature variables selection: In order to reduce the prognostic variables, and improve the prediction accuracy of the model, it is required to carry out high correlation prognostic factors. After One-Way ANOVA analysis, the initial inclusion variables were (P<0.05): age at the time of diagnosis, CS tumor size, histological grade, tumor staging, CS extension , CS lymph nodes, regional nodes positive, marital status, race recode, CS mets at dx, operation type, and whether radiotherapy. Based on One-Way ANOVA analysis and Logistic regression analysis, the prognostic variables were selected as follow (P<0.05): age at diagnosis, tumor size, histological grade, tumor staging, Regional nodes examined, Regional nodes positive. The results are shown in table 2.

5) Data preprocessing: Interval method was selected for metric data to discretize the serious deleted data, false record, and patient's information of deaths from non-lung cancer. The discretization method aimed to divide the interval $[X_0, X_{N-1}]$ into the same subinterval of size D and the suggestion of discretization was given according to the sub-interval index, the observation index *i* and the discrete level *j* satisfy the following conditions [16]:

$$X_0 + \frac{j(X_{N-1} - X_0)}{D} < X_i \leqslant X_0 + \frac{(j+1)(X_{N-1} - X_0)}{D} \quad \cdots\cdots\cdots\cdots (3)$$

The above data preprocessing step is realized by calling bnlearn's function package in software R Studio. Then the data is divided into training set (N1=495) and test set (N2=188) according to the proportion of about 70% and 30% [17]. The training set is used for network learning and adjustment, thus constructing the prognostic model and measuring. The test set is used to evaluate the performance of the model.

6) Construction and prediction of the prognostic model: Bayesian Network (BN) describes the dependent relation of child nodes and the parent nodes by means of the nodes representing the variables and the lines representing the relationships between variables [18], known random variable X = {X1, X2,... Xn}, the distribution of joint probability is:

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \quad \cdots\cdots\cdots\cdots (4)$$

Among them, Pa (Xi) is a subset of Xi parent nodes, and Xi is independent of its non-linear node variables in network graph. Tabu Search (TS) method is used to study bayesian network. This method was proposed by Fred Glover, academician of the American college of engineering in 1986, which is a heuristic algorithm based on neighborhood and iteration to solve and optimize problems. The essence of this method is to prohibit repeating the previous work, jump out of the most advantages of local search, namely random motion and cause a new project in the area, and then evaluate each adjacent solution, and select which can most improve the path of the objective function. If no scheme can improve the results, then a solution with minimum impact on the objective function was chosen,

and find out the best results by imitating human memory [19], steps are as follows:

1) To determine the region N(x), select an initial feasible solution X^0 to make the current optimal solution $X^{best}=X^0$, so $T=N(X^{best})$;

2) According to the above steps, the latest solution X^{next_n}, $n \in [1, +\infty]$is obtained and the output is calculated;

3) Compare the whole decision results and output the optimal solution $X^{next_{max}}=X^{best}$.

The Bayesian prognostic model constructed by Makond et al. [20] was not studied based on the data obtained, but was established by doctors' opinion. The survival model of patients is actually based on empirical modeling. The present study can overcome the disadvantages of single modeling with practical experience; combine the network learning methods with patient's prognosis, and doctor's advice to set up patient's prognosis model. This BN model will be modified and optimized by the R Studio.The network model is shown in Figure 2.

In this study, 188 samples were collected, and 137 were correctly predicted. The prediction accuracy reached 72.87%.

B. Comparison experiment

The decision tree, support vector machine and artificial neural network method are used to establish the prognostic model, and compared the prediction results with prognostic model, which is set up at this study.

In WEKA, the corresponding method of J48, SMO and Multilayer Perceptron are selected to establish the prognostic model and the parameter acquiescence. The prediction accuracy and model performance evaluation of 4 machine learning algorithms are shown in table 3 and table 4.

.

C. Experimental analysis

The study found that the NSCLC prognosis model constructed by Bayesian network was optimal. Table 3 showed that although the decision tree, support vector machine (SVM), and artificial neural network in the predictive accuracy of the training set was higher than that of Bayesian network, but in the testing set three prediction accuracy were significantly reduced in the value compared to the training set, failed to adapt to the new data very well. It is indicated that they were not suitable for practical application, the model of Bayesian networks is better than the other three models in fitting degree. Table 4 also showed that the data in predicting accuracy, accuracy and area under ROC curve of Bayesian network model is significantly higher than that of the other three machine learning algorithms.

Table 1 Prognostic indicator of non-small cell lung cancer patients

Data type	Variables appeared in SEER	Range of value
Categorical variables	Sex	2
	Race recode (Asian)	8
	Marital status at diagnosis	4
	Primary Site – labeled	5
	ICD-O-3 Hist/behav, malignant	4
	Grade	4
	Laterality	2
	CS extension	18
	CS lymph nodes	5
	CS mets at dx	5
	Derived AJCC Stage Group	7
	RX Summ—Surg Prim Site	13
	Radiation	3
Continuous variables	Age at diagnosis	26-90
	CS tumor size	4-132
	Regional nodes positive	0-23
	Regional nodes examined	1-45

Table 2 Screen results of variables based on the Logistic regression analysis

Variables.	B.	S.E.	Exp (B)	95% Exp (B)		Sig.
				Lower limit	Upper limit	
age at diagnosis	-0.066	0.011	0.936	0.916	0.957	0.000
tumor size	-0.018	0.007	0.982	0.968	0.996	0.014
histological grade	/	/	/	/	/	0.001
tumor staging	/	/	/	/	/	0.013
Regional nodes examined	0.050	0.017	1.051	1.016	1.087	0.004
Regional nodes positive	-0.199	0.067	0.819	0.719	0.934	0.003

Table 3 Accuracy predicting among BN-NSCLC model and three other models based on classification algorithms

Classification algorithm	Accuracy predicting	
	Training set	Testing set
Bayesian network	0.683	0.729
Decision tree	0.713	0.670
Support vector machine	0.733	0.686
Artificial neural network	0.784	0.649

Table 4 Performance comparison among models based on different classification algorithms

Classification algorithm	Accuracy predicting	Accuracy	area under the curve of ROC
Bayesian network	72.87%	71.0%	0.67
Decision tree	67.02%	66.3%	0.568
Support vector machine	68.62%	68.2%	0.611
Artificial neural network	64.89%	63.7%	0.615

The choice of network learning method is the basis of constructing Bayesian classifier. This study selects the Testing set (TS) method to build a network model, preliminary is the optimization of climbing method. When the known variables do not constitute a network, mobile search could instead of random generating, and using of three actions of addition, subtraction and the reverse side to create neighborhood, and search the global optimal solution to adjust the structure of the network to complete the self learning of Bayesian networks. On this basis, combining with the experience of clinical doctors to modify the network graph. Making the high correlation prognostic factors associated is a typical combination of theoretical method and practical application.

The adjustment of network graph is the most critical process for the construction of the survival prediction model. As shown in Figure 2, the arrow direction indicated the relations between nodes, such as the size to stage indicate the direct impact on the latter, the former selected prognostic variables are pointing to the final survival state. The age at diagnosis, tumor stage, and affected lymph node ratio directly affect the survival of patients. By constructing different network diagram to find the optimal classification model, so as to determine the relationship between the prognostic factors and the impact on the survival, and accordingly, the conditions of prognosis could be evaluated and the related factors also could be controlled. Of course, because of the used SEER database in present study did not adopt all tumor prognostic factors [21], so the numbers of selected variables in the model are limited, and the prediction model maybe exist some certain limitations.

V. CONCLUSION

In this study, the survival prognosis model of patients with non-small cell lung cancer patients was established, and the prediction accuracy was 72.87%. By constructing a Bayesian network to explore the relationship between the prognostic variables and the impact on the survival situation, and on the basis of the internal network structure adjustment combined with clinical expert advice, the relationship between nodes in the model was better explained. Firstly Apply the SEER database in Asia cancer as the main research object to build its prediction model, and playing a supplementary role on the judgment of the prognosis of patients with postoperative five years. In future research, the inclusion of external validation of other sources can be considered, and the adaptation degree of prediction model itself can be improved and better serve for clinical treatment and prognostic evaluation.

REFERENCES

[1] L. Ngo, P. Haddawy, R.A. Krieger, and J. Helwig, "Efficient temporal probabilistic reasoning via context-sensitive model construction," *Comput. Bio. Med.*, vol.27, no.5, pp. 453-476, 1997

[2] P.J. Lucas, L.C. Van, and A. AbuHanna, "Bayesian networks in biomedicine and health-care," *Artif. Intell. Med*., vol. 30, no.3, pp. 201-214, 2004

[3] A. Dekker, C. Dehingoberije, D.D. Ruysscher, K. Komati, G. Fung, S. Yu, P.A. Malvern, A. Hope, and W.D. Neve, "Survival prediction in lung cancer treated with radiotherapy: Bayesian networks vs. support vector machines in handling missing data," *Proceedings of the 2009 International Conference on Machine Learning and Applications* , pp. 494-497, 2009

[4] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, and A.L.A.J. Dekkera, "Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy," *Med. Phys.,* vol. 37, pp. 1401-1407, 2010.

[5] M. Velikova, M. Samulski, P.J.F. Lucas, and N. Karssemeijer, "Improved mammographic CAD performance using multi-view information: a Bayesian network framework," *Phys. Med. Biol.,* vol. 54, pp. 1131-1147,2009.

[6] X. Chen, G. Anantha, and X. Wang, "An effective structure learning method for constructing gene networks," *Bioinformatics*, vol. 22, pp. 1367-1374, 2006.

[7] M.A. van Gerven, B.G. Taal, and P.J. Lucas, "Dynamic Bayesian networks as prognostic models for clinical patient management," *J. Biomed. Inform.*, vol. 41, pp. 515-529, 2008.

[8] W. Smith, J. Doctor, J. Meyer, I. Kalet, and M. Philips, "A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model," *Artif. Intell. Med*., vol. 46, pp. 119-130, 2009

[9] American cancer society: Cancer facts and figures. Atlanta, GA:American Cancer Society, 2008.

[10] National Cancer Institute. SEER Cancer Statistics Review (CSR) 1975-2013[R/OL].[2016-09-20].http://seer.cancer. gov/csr/1975_2013/sections.html

[11] J.G. Armstrong, M.J. Zelefsky, S.A. Leibel, C. Burman, C. Han, L.B.Harrison, G.J. Kutcher, and Z.Y. Fuks, "Strategy for dose escalation using 3-dimensional conformal radiation therapy for lung cancer," *Ann. Oncol*., vol. 6, pp. 693-697, 1995.

[12] A. Abramyuk, S. Tokalov, K. Zophel, A. Koch, K. Szluha Lazanyi, C. ¨ Gillham, T. Herrmann, and N. Abolmaali, "Is pre-therapeutical FDGPET/CT capable to detect high risk tumor subvolumes responsible for local failure in non-small cell lung cancer?," *Radiother. Oncol*., vol. 91,pp. 399-404, 2009.

[13] Shin H, Nam Y, "A coupling approach of a predictor and a descriptor for breast cancer prognosis," *BMC Medical Genomics*, vol. 7, pp. S1-S4, 2014.

[14] American joint committee on cancer, AJCC cancer staging manual [M]. The 7th Edition. *New York: Springer Verlag*, pp. 253-270, 2010.

[15] 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.

[16] A.J. Hartemink, "Principled computational methods for the validation and discovery of genetic regulatory networks," *Massachusetts Institute of Technology*, pp. 86-87, 2001.

[17] Y. Kumar, G. Sahoo, "Prediction of different types of liver diseases using rule based classification model" *Technology & Health Care Official Journal of the European Society for Engineering & Medicine,* vol. 21, pp. 417-432, 2013.

[18] J.H. Oh, J. Craft, L.R. Al, M. Vaidya, Y. Meng, "A Bayesian network approach for modeling local failure in lung cancer," *Physics in Medicine & Biology*, vol. 56, pp. 1635-1651, 2011.

[19] W.L. Lim, A. Wibowo, M.I. Desa, H. Haron, "A biogeography-based optimization algorithm hybridized with tabu search for the quadratic assignment problem," *Computational Intelligence & Neuroscience*, vol. 2016, pp. 27-38, 2016.

[20] B. Makond, K.J. Wang, K.M. Wang, "Probabilistic Modeling of Short Survivability in Patients with Brain Metastasis from Lung Cancer," *Computer Methods & Programs in Biomedicine*, vol. 119, pp. 142-162, 2015.

[21] Q. Yang, J.P. Zhang, "Clinical Applications of the Tumor Registry Database," *The Journal of Evidence-Based Medicine*, vol. 13, pp. 250-251, 256, 2013.