Prediction of Survival in Surgical Unresectable Lung Cancer by Artificial Neural Networks Including Genetic Polymorphisms and Clinical Parameters

Te-Chun Hsia,[1] Hung-Chih Chiang,[2,3] David Chiang,[2] Liang-Wen Hang,[1] Fuu-Jen Tsai,[4,5] and Wen-Chi Chen[4,6]*

[1]Department of Internal Medicine, China Medical College Hospital, Taichung, Taiwan
[2]Department of Management, National Taiwan University, Taipei, Taiwan
[3]Ching Yun Institute of Technology, Chungli
[4]Department of Medical Genetics, China Medical College Hospital, Taichung, Taiwan
[5]Department of Pediatrics, China Medical College Hospital, Taichung, Taiwan
[6]Department of Urology, China Medical College Hospital, Taichung, Taiwan

Lung cancer, a common malignancy in Taiwan, involves multiple factors, including genetics and environmental factors. The survival time is very short once cancer is diagnosed as being in advanced stage and surgically unresectable. Therefore, a good model of prediction of disease outcome is important for a treatment plan. We investigated the survival time in advanced lung cancer by using computer science from the genetic polymorphism of the p21 and p53 genes in conjunction with patients' general data. We studied 75 advanced and surgical unresectable lung cancer patients. The prediction of survival time was made by comparing real data obtained from follow-up periods with data generated by an artificial neural network (ANN). The most important input variable was the clinical staging of lung cancer patients. The second and third most important variables were pathological type and responsiveness to treatment, respectively. There were 25 neurons in the input layer, four neurons in the hidden layer-1, and one neuron in the output layer. The predicted accuracy was 86.2%. The average survival time was 12.44 ± 7.95 months according to real data and 13.16 ± 1.77 months based on the ANN results. ANN provides good prediction results when clinical parameters and genetic polymorphisms are considered in the model. It is possible to use computer science to integrate the genetic polymorphisms and clinical parameters in the prediction of disease outcome. Data mining provides a promising approach to the study of genetic markers for advanced lung cancer. J. Clin. Lab. Anal. 17:229–234, 2003. © 2003 Wiley-Liss, Inc.

Key words: artificial neural networks; data mining; lung cancer; p21 gene; p53 gene; survival

INTRODUCTION

Lung cancer is a multifactorial and complex disease and the molecular mechanisms initiating disease progression and prognosis are not well understood. According to pathology and natural history, lung cancer can be divided into two major groups: small cell lung cancer (SCLC) and non-small cell lung carcinoma (NSCLC), the latter representing 80% of patients with lung cancer in Taiwan (1). Most of the patients with NSCLC are inoperable and already in advanced stages (stage IIIB or IV) at diagnosis. Therefore, the prediction of disease outcome is indeed important for determining further therapy.

Computer science has been used as an analytic tool by researchers in several medical fields. The binary classification problem has wide applications to problems in biology and medical domains. Data mining tools, such as artificial neural networks (ANNs), have been

Grant sponsor: China Medical College Hospital; Grant number: DMR-92-069.

*Correspondence to: Wen-Chi Chen, Department of Urology, China Medical College Hospital, No. 2, Yu-Der Road, Taichung, 404, Taiwan. E-mail: drtom@www.cmch.org.tw

Received 17 April 2003; Accepted 12 June 2003
DOI 10.1002/jcla.10102
Published online in Wiley InterScience (www.interscience.wiley.com).

© 2003 Wiley-Liss, Inc.

well documented and are also used for the prediction of diseases, treatment outcomes, and prognosis of a variety of diseases. Various architectures of ANN have been used in different medical diagnoses and their results have been compared with existing classification methods and physicians' diagnoses (2). ANNs are simple models of the way the biological nervous system operates and the network is able to learn through training. As training progresses, the network usually becomes much more accurate in replicating the known outcomes. Once we have trained the neural network, it can be applied to future cases where the outcome is known. Lin et al. (3) first applied ANNs as a false–positive detection tool in digital chest radiographs for the diagnosis of lung cancer. Since that time, ANNs have been used widely in the reduction of false–positive detection of chest nodules from chest digital radiographs (4). However, few studies regarding censored survival data of lung cancer by using ANNs have been reported (5).

Recently, single nucleotide polymorphisms (SNPs) have been used by researchers as a tool in the search for genetic variations and the environmental influence on many complex diseases (6). In this study, the genetic polymorphisms that were thought to be associated with lung cancer were p53 gene codon 72 polymorphism and p21 gene codon 31 polymorphism (7–9). There have been no previous studies of ANNs using genetic polymorphisms in conjunction with clinical parameters as input to predict outcome of lung cancer. The results of this study provide a new method for the study of the prognosis in advanced lung cancer.

MATERIALS AND METHODS

Patient Selection

A total of 76 patients (54 males and 21 females), between the ages of 32 and 84 years (mean: 63.9 years) with advanced lung cancer diagnosed at China Medical College Hospital, Taiwan, were enrolled in this study. The diagnosis was confirmed by throacotomic biopsy, bronchoscopic brushing and washing, or echo-guided fine needle aspiration. The patients were at least stage II and advanced NSCLC in pathological grading according to the TNM system at initial diagnosis (10,11). Informed consent was obtained from both groups participating in this study.

Polymerase Chain Reaction (PCR)

We performed two PCRs for the p53 gene and p21 gene polymorphisms in each patient. PCRs were carried out in a total volume of 50 μl, containing genomic DNA; 2–6 *p*mole of each primer; 1X Taq polymerase buffer (1.5 mM $MgCl_2$); and 0.25 units of AmpliTaq DNA polymerase (Perkin Elmer, Foster City, CA). Primers for Arg72 for p53 codon 72 were: 5′-TCCCCC-TTGCCGTCCCAA-3′ and 5′-CTGGTGCAGGGGC-CACGC-3′ according to Storey et al. (12). Primers of Pro72 for p53 codon 72 were: 5′-GCCAGAGGCTGC-TCCCCC-3′, and 5′-CGTGCAAGT CACAGACTT-3′. The primer pairs for p21 codon 31 were designed from codon 1 to codon 91 (5′-GTCAGAAC CGGCTGGG-GATG-3′ and 5′-CTCCTCCCAACTCAT CCCGG-3′), according to the procedure described by Li et al. (13).

The PCR products of arg72 and pro72 from the same individual were mixed together and 10 μl of the resulting solution was loaded into a 3% agarose gel containing ethidium bromide for electrophoresis to detect the p53 gene codon 72 polymorphism. PCR resulted in two bands of 177-bp and 141-bp representing proline and arginine, respectively. One band of PCR product in agarose gel electrophoresis represented homozygotes and two bands were heterozygotes. The PCR product of 272-bp of the p21 gene was mixed with two units *Blp* I (New England Biolabs, Beverly, MA) and reaction buffer according to the manufacturer's instruction. A restriction site was located at the frequently seen allele (AGC) of codon 31 to form an excisable site. Two fragments of 89-bp and 183-bp were present if the product was excisable.

Chemotherapy and Evaluation of Results

At every cycle of chemotherapy, patients received gemcitabine (Lilly, Indianapolis, Indiana) at a dose of 800–1,200 mg/m^2 on day 1 and day 8 and received gemcitabine at a dose of 800–1,200 mg/m^2 plus cisplatin (Lilly, Indianapolis, Indiana) 100 mg/m^2 on day 15 (14). After a cycle of chemotherapy, patients took a 2 week rest. Patients who did not complete 4–6 cycles of chemotherapy were not included. The treatment response was recorded according to the World Health Organization (WHO) criteria for the assessment of efficacy by chemotherapy (10). A complete response was defined as the complete disappearance of all evidence of tumors. Partial response was defined as a greater than 50% reduction in the sum of the products of the largest perpendicular diameters of all measured lesions for at least 4 weeks. Stable disease was defined as a decrease of less than 50% or an increase of less than 25% of well-outlined lesions for at least 4 weeks. Progressive disease was defined as an increase of greater than 25% of the cross-sectional area of one or more lesions, or the occurrence of a new lesion (15). Patients were then subdivided into two groups according to their response to chemotherapy. Patients with a complete response or a partial response were categorized as group 1, whereas the others were categorized as group 2 (nonresponse).

Computing Techniques

We entered all patient and genetic polymorphism data into a computer program in order to progress the ANN. Data from each patient were: sex, age, smoking habit, tumor staging, cell type, and responsiveness to chemotherapy. The final results of the patient's survival period were used as output in ANN. All calculations regarding ANN were performed with the SPSS® (Statistical Packages for the Social Science, Chicago, IL) Clemetine data mining software (16). The predicted value was compared to the clinical survival data and the rate of accuracy was defined as:

$$\text{Rate of accuracy (\%)} = (1 - \text{Sum}\,(|A - B|)\,/\,\text{Sum}\,(A))\times 100$$

where A is the survival time of each patient and B is the survival time of the ANN prediction.

A sensitivity analysis of the input parameters was performed after the ANN network was trained. It provides information on which input parameters were most important in predicting the output parameters (survival years for the patient). The analysis was based on the classification function of ANN. Each input parameter has relative importance during process.

According to the pathological cell types, there were 19 squamous cell carcinoma, five small cell carcinoma, and 51 adenocarcinoma patients. Clinical staging included six patients with IIa, three with IIb, seven with IIIA, 18 with IIIB, and 41 with stage IV. Of the 76 patients, 37 patients had a smoking habit. There were 12 smokers among the 19 squamous cell carcinoma patients (63.1%) and four smokers among the five small cell carcinoma patients (80%). The remaining 21 smokers had adenocarcinoma (41.2%). PCR results of the p53 and p21 gene polymorphisms are listed in Table 1.

RESULTS

This study used all the input variables listed in Table 2 as the input vectors for the ANNs. The suitability and performance of an appropriate network or learning method and the corresponding combination of these variables were iteratively analyzed. The network configuration and additional information about the learning algorithm is provided in Table 3. This study tried different combinations of hidden layers and found out that one hidden layer yielded the best result for our data. The other related configuration parameters that

TABLE 1. Distribution of p53 codon 72 polymorphisms and p21 codon 31 polymorphism in healthy control subjects and lung cancer patients

P53 gene	Pro/Pro	Pro/Arg	Arg/Arg	Total	χ^2	*P* value
Control	8 (13.5%)	26 (44.0%)	25 (42.4%)	59 (100%)		0.008[a]
Lung cancer	15 (31.9%)	24 (51.1%)	8 (17.0%)	47 (100%)	9.734	
P21 gene	CC	C/A	AA	Total	χ^2	*P* value
Control	21 (17.6%)	60 (50.4%)	38 (31.9%)	119 (100%)		0.099
Lung cancer	8 (17.0%)	16 (34.0%)	23 (48.9%)	47 (100%)	4.63	

[a] $P < 0.01$.

CC, serine homozygote; AA, arginine homozygote; C/A, heterozygote.

TABLE 2. Study variables

Number	Variables	Description	In/output variables
1.	Gender	Patient's sex	Input
2.	Age	Patient's current age	Input
3.	Gene (p21)	Gene type of p21	Input
4.	Gene (p53)	Gene type of p53	Input
5.	Disease	Disease type of patient's lung cancer	Input
6.	Period of lung cancer	Period of patient's lung cancer	Input
7.	Chemical diagnosis	Treatment type of chemical diagnosis	Input
8.	Smoking habit	Patient's smoking habit	Input
9.	Prediction of survival	Prediction of patient's survival years	Output

TABLE 3. Parameters and configuration of the applied artificial neural network

Parameter	Value
Learning mode type	Prune method
Prevent overtraining	Yes
Training percent (%)	50
Sensitivity analysis	Yes
Stop on	Default
Set random seed	No
Generate model from	Best network
Hidden layers	1
Hidden units (#1 hidden layer)	20, 15
Hidden rate	0.15
Input rate	0.15
Hidden persistence	6
Input persistence	4
Persistence	100
Overall persistence	3
Learning rate initial elta	0.3
Eta range	[0.01,0.1]
Momentum term alpha	0.9

TABLE 4. Sensitivity analysis of the input parameters for the most successful calculation with an artificial neural network (ANN)

Priority	Input variables	Relative importance
1	Staging	0.04418
2	Cell type	0.04075
3	Response to chemotherapy	0.04013
4	Gene (p53)	0.03197
5	Smoking	0.02640
6	Sex	0.02436
7	Gene (p21)	0.02099
8	Age	0.00592

yielded good result for our ANN model are listed in Table 3.

Details of the sensitivity analysis for the ANN network are given in Table 4. The most important input variable was clinical staging of lung cancer. The second and third most important variables were patient's pathological cell type of lung cancer and response to chemotherapy, respectively.

The architecture of ANN in this prediction model has 25 neurons at the input layer, four neurons in the hidden layer-1, and one neuron at the output layer. The predicted accuracy was 86.2% by ANN. The average survival time was 12.44 ± 7.95 (range 4–34) months in real data and 13.16 ± 1.77 (range 3.3–13.6) months from ANN results. The relative importance of the input parameters is shown in Table 4 in descending order. Tumor staging was the most important parameter in this prediction model and age was the least important in ANNs prediction. The p53 gene polymorphism contributed to the fourth degree of importance and the p21 gene polymorphism was contributed to the seventh degree of importance.

DISCUSSION

ANN achieved promising classification results when clinical parameters and genetic factors were considered simultaneously in the prediction model. ANN had a prediction success rate of 86.2%. The prediction model can be a useful method to predict, with a high success rate, the clinical outcome of advanced lung cancer. Although the success rate of correct prediction was not 100%, this study shows that the rate can be improved step by step when parameters and genetic factors involved in lung cancer are considered together. This study was able to show each factor's importance priority in lung cancer by observing the fractional percentages. By using these methods, the hidden meanings behind the patients' data can be further revealed, providing much more accurate predictions of outcome in lung cancer. The ANN prediction model has the capability to deal with different kinds of medical information, such as model creation, model learning, problem classification, generalization, and data interpretation. ANN has been used primarily in applications such as heart disease diagnosis, skin diseases diagnosis, and headache treatments (17–19). ANN has been used less in predicting the survival time for lung cancer patients in Taiwan. To our knowledge, this is the first preliminary report.

ANNs have their limitations and problems. One of the limitations is the way a trained network makes its decisions. Because the information encoded by the network is just a collection of numbers, it is quite difficult to work out the reasoning that goes into its decision-making process. Neural networks are sometimes referred to as black boxes because of this limitation. In order to solve this kind of problem, rule induction technique might be the answer. Rule induction, also known as decision tree technique, is the complimentary technique to neural network. Working from either the complete data set or a subset, rule induction creates a decision tree, representing a rule for how to classify the data into different outcomes. The tree's structure, and hence the rule's reasoning process, is open and explicit and can be browsed. In order to gain many implications from the prediction model, our future work will focus on using the decision tree technique together with neural network.

Another problem when using neural network is how to configure the network topology and appropriately control parameters. The connections between neurons in the neural network are very complicated and the configuration parameters of the network require much manual trial and error in order to gain an appropriate and functional network. The solution to this problem might be to use genetic algorithms (GA) to pick out the appropriate configuration parameters for the neural network (20).

There was a relative short range of survival of ANN's result when compared with the real data. This difference was caused by the computer program exerting a fitness function to predict the survival, instead of using real data acquired in every patient's data training. The standard deviation of survival data from ANN's result was also less than that from the real data. In general, ANN combines the program's functions (function of combination) to predict the fittest overall result that makes a shorter difference among output in each patient. Therefore, a small standard deviation was obtained by ANN's prediction.

Traditionally, tumor staging, performance status, and body weight loss were used as prognostic factors for lung cancer. Staging, cell type, responsiveness to chemotherapy, and the p53 gene polymorphism were important factors for the prediction of patient survival in this study. Our results indicated that staging is the most important factor. Although polymorphisms of the p53 and p21 genes were reported to be associated with pathogenesis of lung cancer, only the p53 gene polymorphism was important in the processing of data. More genetic data may be helpful in improving the prediction accuracy.

There was a short survival period in patients with surgically unresectable lung cancer. This may be due to late diagnosis or poor response to chemotherapy. ANNs can help in the diagnosis of lung cancer from suspected nodule in the chest radiography or computerized tomography and by analyzing specimens of fine needle biopsy (3,4, 21,22). In conjunction with our study, it is predicted that data mining such as ANN will be a powerful tool and will be applied in all fields of medical science—from disease diagnosis to outcome prediction and management. Therefore, ANNs will be helpful in improving diagnosis and management of short survival time in lung cancer patients.

In summary, ANN yielded good prediction results when clinical parameters and genetic polymorphisms in the model were considered. ANN was able to reach a successful classification rate of over 86%. Using computer science to integrate genetic polymorphisms and clinical parameters in the prediction of disease outcome seems promising. Even though parameter configurations and modifications need further trials when running the final results, it is still worthwhile to setup an ANN in order to acquire better prediction results for the model. Data mining provides a new approach to the study of genetic markers for far-advanced lung cancer. This model is a new modality in the study of lung cancer, complementing the study of genetic polymorphisms and clinical parameters.

REFERENCES

1. Department of Health, the Executive Yuan. Republic of China, general health statistics, 1997. In: Health and vital statistics. Taipei, Taiwan: R.O.C. Press; 1998. p 86–108.
2. Finne P, Finne R, Stenman UH. Neural network analysis of clinicopathological factors in urological disease: a critical evaluation of available techniques. BJU Int 2001;88:825–831.
3. Lin JS, Ligomenides PA, Freedman MT, et al. Application of artificial neural networks for reduction of false-positive detections in digital chest radiographs. Proc Annu Symp Comput Appl Med Care 1993;434–438.
4. Wu YC, Doi K, Giger ML, et al. Reduction of false positives in computerized detection of lung nodules in chest radiographs using artificial neural networks, discriminant analysis, and a rule-based scheme. J Digit Imaging 1994;7:196–207.
5. Biganzoli E, Boracchi P, Mariani L, et al. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Stat Med 1998;17:1169–1186.
6. Loder N. Genetic variations can point the way to disease gene. Nature 1999;401:734.
7. Shih CM, Lin PT, Wang HC, et al. Lack of evidence of association of p21 WAF1/CIP1 polymorphism with lung cancer susceptibility and prognosis in Taiwan. JPN J Cancer Res 2000;91:9–15.
8. Wang YC, Chen CY, Chen SK, et al. p53 codon 72 polymorphism in Taiwanese lung cancer patients: association with lung cancer susceptibility and prognosis. Clin Cancer Res 1999;5:129–134.
9. Wang YC, Lee HS, Chen SK, et al. Prognostic significance of p53 codon 72 polymorphism in lung carcinomas. Eur J Cancer 1999;35:226–230.
10. Anonymous. The world health organization. Histological typing of lung tumors. Neoplasia 1982;29:111–123.
11. Moutain CF. Revisions in the international system for staging lung cancer. Chest 1997;111:1710–1717.
12. Storey A, Thomas M, Kalita A, et al. Role of a p53 polymorphism in the development of human papillomavirus-associated cancer. Nature 1998;393:229–234.
13. Li YJ, Laurent-Puig P, Salmon RJ, et al. Polymorphisms and probable lack of mutation in the WAF1-CIP1 gene in colorectal cancer. Oncogene 1995;10:599–601.
14. Abratt RP, Sandler A, Crino L. Combined cisplatin and gemcitabine for NSCLC: influence of scheduling on toxicity and drug delivery. Semin Oncol 1998;25(suppl):35–43.
15. Treat J. Part II: treatment of non-small cell lung cancer: chemotherapy. In: Fishman AP, Elias JA, Fishman JA, et al., editors. Fishman's pulmonary diseases and disorders, Vol. II, 3rd ed. New York: McGraw-Hill 1998; p 1797–1818.
16. SPSS Data Mining. 1998. http://www.spss.com/datamine/whitpap.htm.
17. Binder M, Steiner A, Schwarz M, et al. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. Br J Dermatol 1994;130: 460–465.
18. de Tommaso M, Sciruicchio V, Bellotti R, et al. Photic driving response in primary headache: diagnostic value tested by

discriminant analysis and artificial neural network classifiers. Ital J Neurol Sci 1999;20:23–28.

19. Mobley BA, Schechter E, Moore WE, et al. Predictions of coronary artery stenosis by artificial neural network. Artif Intell Med 2000;18:187–203.
20. Cho SJ, Hermsmeier MA. Genetic algorithm guided selection: variable selection and subset selection. J Chem Inf Comput Sci 2002;42:927–936.
21. Matsuki Y, Nakamura K, Watanabe H, et al. Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high resolution CT: evaluation with receiver operating characteristics analysis. AJR Am J Roentgenol 2002;178:657–663.
22. Zhou ZH, Jiang Y, Yang YB, et al. Lung cancer cell identification based on artificial neural network ensembles. Artif Intell Med 2002;24:25–36.