35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Predicting Post-Treatment Survivability of Patients with Breast Cancer Using Artificial Neural Network Methods

Tan-Nai Wang, Chung-Hao Cheng, Hung-Wen Chiu, *Member, IEEE*

Abstract—**In the last decade, the use of data mining techniques has become widely accepted in medical applications, especially in predicting cancer patients' survival. In this study, we attempted to train an Artificial Neural Network (ANN) to predict the patients' five-year survivability. Breast cancer patients who were diagnosed and received standard treatment in one hospital during 2000 to 2003 in Taiwan were collected for train and test the ANN. There were 604 patients in this dataset excluding died not in breast cancer. Among them 140 patients died within five years after their first radiotherapy treatment.**

The artificial neural networks were created by STATISTICA® software. Five variables (age, surgery and radiotherapy type, tumor size, regional lymph nodes, distant metastasis) were selected as the input features for ANN to predict the five-year survivability of breast cancer patients. We trained 100 artificial neural networks and chose the best one to analyze. The accuracy rate is 85% and area under the receiver operating characteristic (ROC) curve is 0.79. It shows that artificial neural network is a good tool to predict the five-year survivability of breast cancer patients.

I. Introduction

Breast cancer is the most common cancer in women worldwide and is diagnosed in more than 1 million women each year. Breast cancer is largely seen in women in economically developed areas such as North America and Northern, Western, and Central Europe, and in Asian metropolises, including Tokyo, Shanghai, Hong Kong, Singapore, and Taiwan. 2012 statistics from the Department of Health, Executive Yuan indicate that breast cancer is the fourth leading cause of death for women in Taiwan and that 1852 women died from breast cancer in 2011 [1]. Therefore, the early detection and treatment of breast cancer is a major public health issue.

In addition to determining the clinical status of patients, cancer staging systems can serve as references for selecting clinical treatment methods and for predicting prognosis. According to the tumor-node-metastasis (TNM) staging system proposed by the American Joint Committee on Cancer in 2002, the stages of breast cancer in patients are determined by tumor size (T), whether the tumors have metastasized to the lymph nodes (N), and whether tumors have metastasized to other organs (M).

In numerous previous studies concerning the prediction of 5-year survival rate for patients suffering from breast cancer, the variables include age, tumor stage, ethnicity, socioeconomic status [2], lymph node status, histological type, tubule formation, tumor necrosis [3], marital status, the degree of tumor lesion, pathological stage, the type of radiation treatment, clinical stage, the degree of tumor spread, the number and size tumors, and the number of lymph nodes [4]. The use of artificial neural networks for predicting the survival rate of patients suffering from breast cancer in Taiwan and foreign countries has yielded satisfactory results [4,5].

Prognostic factors in cancer have great effects on clinical decision-making and they provide patients with greater health-related information. Therefore, we used the database of breast cancer patients at a regional hospital in Northern Taiwan as examples to establish a prediction model for predicting the 5-year survival rate of patients suffering from breast cancer following treatment. In addition, we analyzed and discussed the outcomes of this model.

II. Methods and Materials

We adopted the database of a regional hospital in Northern Taiwan that had 5190 patients with breast cancer between 2000 and 2003. The research targets were female patients who were preliminarily diagnosed with breast cancer and received their entire treatment in the study hospital. Incomplete data sets and the data of patients who died of causes unrelated to breast cancer were eliminated from the analysis. Consequently, 604 comprehensive patient data were available for prediction analysis. Among these 604 patient data, 141 patients died within 5 years, which accounted for 23.34% of the total population of patients suffering from breast cancer.

After organizing the 604 data sets, five variables were selected for establishing an artificial neural network model using STATISTICA® software to predict the 5-year survival rate of the patients. The five variables selected were age (A), tumor size (T), condition of tumor metastasizing to lymph nodes (N), whether the tumors had metastasized to other organs (M), and whether breast-conserving surgery was performed followed by radiotherapy. The feature selection was based on literature review and clinical availability. We attempt to build a model only using data at hand.

Classification was selected as the prediction method for the establishment of the artificial neural network model. The survival variable of the patients surviving beyond 5 years was set as 1. The survival variable of patients who died after 5

*Research supported by National Science Council, Taiwan.

T-N Wang is with the Graduate Institute of Biomedical Informatics and Office of Biomedical Informatics, Taipei Medical University, Taipei 110, Taiwan (e-mail: dance@tmu.edu.tw).

C-H Cheng is with the Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei 110, Taiwan. He is also now with the Department of Radiation Oncology, Koo Foundation Sun Yat-Sen Cancer Center, New Taipei City, Taiwan (e-mail: mac.cheng@msa.hinet.net).

H-W Chiu is with the Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei 110, Taiwan (Corresponding author to provide phone: 886-2-27361661; fax: 886-2-27339049; e-mail: hwchiu@tmu.edu.tw).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

years was established as 0. The breast-conserving variable of patients who received breast-conserving surgery followed by radiotherapy that delivered a dose greater than 6200 cGy was set as 1. The breast-conserving variable for patients who received breast-conserving surgery followed by radiotherapy that delivered a dose smaller than 5000 cGy was set as 0.

To prevent over-training from occurring, 75% of the samples in the database were randomly selected as training samples using the software and the remaining 25% were used as test samples. Among the five variables, patient age at diagnosis was used as a continuous variable; whereas tumor size, condition of tumor metastasizing to lymph nodes, whether the tumors had metastasized to other organs, and whether breast-conserving surgery was performed followed by radiotherapy were used as class variables. The artificial neural network used in this study was a multilayer perceptron (MLP) artificial neural network and the number of hidden neuron was set as ranging from 5 to 15. We applied the error evaluation method of minimum cross entropy to train 100 artificial neural networks and studied and analyzed the network that had the best overall performance.

III. Results

The data used in this study contained 604 data with comprehensive patient information. In the classification based on the final survival status, 141 patients died within 5 years, accounting for 23.34% of the total patient population, and 563 patients survived beyond 5 years, accounting for 76.66% of patient population. In the classification based on cancer stage, the number of patients diagnosed as in Stage I to Stage IV were 172, 188, 218, and 26, respectively, accounting for 28.48%, 31.13%, 36.09%, and 4.30% of the total patient population, respectively. In the classification based on whether the patients received breast-conserving surgery and radiotherapy, the number of patients who received breast-conserving surgery followed by radiotherapy with a delivered dose greater than 6200 cGy was 296, 49% of the patient population, and the number of patients who received breast-conserving surgery followed by radiotherapy that delivered a dose smaller than 5000 cGy was 308, accounting for 51% of the population. Among the total patient population, 453 patients (75%) were randomly selected by computer software as members of the training group, and the remaining 151 patients (25%) were grouped into the test group.

The optimized model established by STATISTICA® software using MLP artificial neural network was MLP 13-6-2. The algorithm used was the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, the training performance was 85.43, and the testing performance was 84.11.

The 5-year survival rate predicted using the artificial neural network was compared to the original data as shown in TABLE I. The accuracy was calculated using a formula as 85.1%, the sensitivity as 94.83%, and the specificity as 52.86%.

The receiver operating characteristic (ROC) curve of the training samples obtained by separately analyzing the training samples and test samples is shown in Fig. 1. The threshold was 0.2262 and the area under the curve was 0.8765. The results suggested that the accuracy of the training samples was excellent.

TABLE I. Survival prediction results

		Actual value		Total
		Survival	Death	
Prediction outcome	Survival	440	66	506
	Death	24	74	98
Total		464	140	604

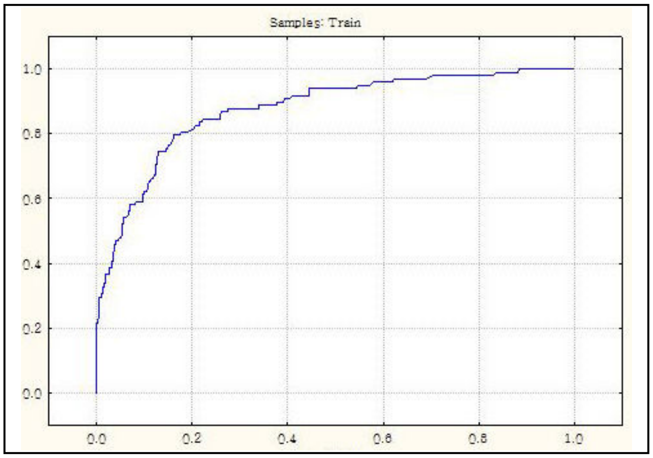

Figure 1. ROC curve of training samples

The ROC curve of the test samples is shown in Fig. 2. The threshold used was 0.14462 and the area under the curve was 0.7935. The results suggested that the accuracy of the test samples was moderate.

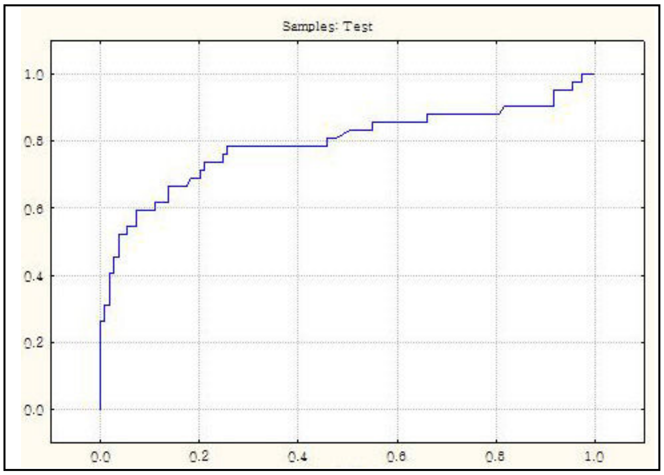

Figure 2. ROC curve of testing samples

IV. Conclusions and Discussion

We generated cancer stage-survival rate correlation curves and analyzed and verified whether the STATISTICA® software genuinely randomly assigned samples to the training and test groups and whether the data were evenly distributed.

By comparing the data shown in Figs. 3 and 4, a differential survival rate between the training and test groups was observed in Stage IV patients; whereas the differences in the distribution of the survival rates were small in the Stage I to III patients. The differential survival rate observed in the Stage IV patients may have been attributed to the small sample size. However, because none of the Stage IV patients survived beyond 5 years, the differential survival rate did not affect the prediction. Thus, the sample assignment was random.

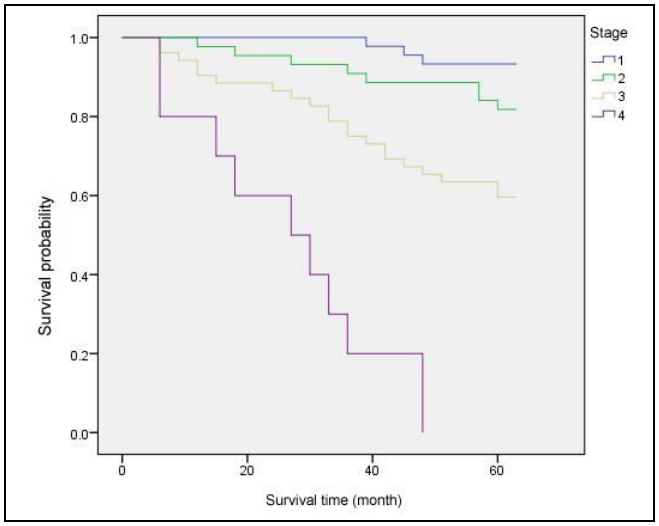

Figure 3. Kaplan-meier survival curve of testing samples

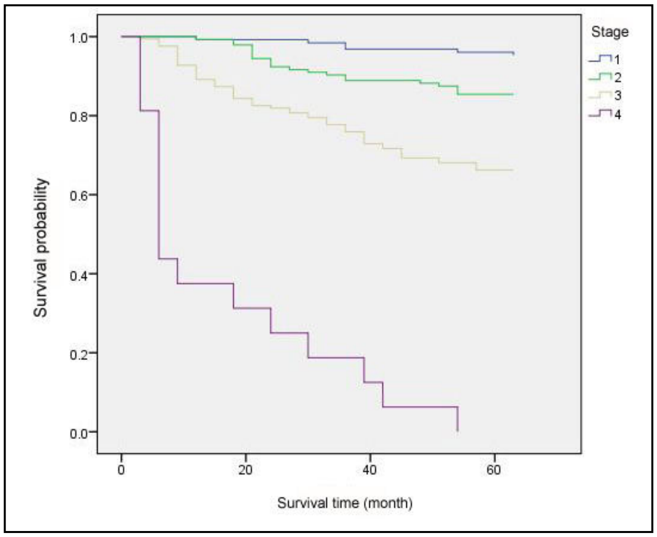

Figure 4. Kaplan-meier survival curve of training samples

The correlation coefficients between various variable entries and survival outcome were shown in Table II; the correlation was significant if the level of significance was determined to be 0.01. As shown in Table II, patient survival had a moderate negative correlation with the five variables (tumor size, condition of tumor metastasizing to lymph nodes, whether tumors have metastasized to other organs, whether breast-conserving surgery was performed, and radiotherapy), a moderate positive correlation with whether patients received breast-conserving surgery, and a low negative correlation with age of diagnosis. However, the correlations of these variables to survivals were not good enough so that it could not get accurate prediction by only one variable.

TABLE II. CORRELATION COEFFICIENT BETWEEN INPUT VALUES AND SURVIVAL

correlation coefficient	Age	T	N	M	Treatment
Survival	-0.172	-0.450	-0.418	-0.386	0.382

In the analysis of various numerical indices, the comparison of the predicted 5-year survival rate using the artificial neural network model established as described in this study and the actual patient survival condition indicated that the accuracy was 85.1%, the sensitivity was 94.83%, and the specificity was 52.86%. As the data imply, the artificial neural network model was highly accurate and sensitive. The reason why the artificial neural network model provided high accuracy and sensitivity was that the thresholds selected for both the training and test samples were low. Consequently, the predicted patient survival rates were inclined to be high, and with the high ratio of surviving patients (76.66%) in these patient data, the model yielded high accuracy and sensitivity. Comparing with other studies, this result is not remarkable. In the study published by Delen et al.[4], the accuracy of prediction reach to 0.9 above. However, they used 17 feature variables to perform such a work, we only used 5 features at hand to get accuracy over 0.85. This simple system will be more suitable for clinical application.

Another noteworthy point in this study was the relatively low specificity (52.86%). We found that the reason for the low specificity was the presence of numerous false positive cases; that is, numerous patients classified in an early stage of breast cancer were predicted to survive beyond 5 years but died within 5 years. In addition to the three cancer-staging variables (i.e., tumor size, condition of tumor metastasizing to lymph nodes, and whether tumors had metastasized to other organs), additional and more crucial malignant factors may have caused the patients in the early stage to die within 5 years. These additional variables were not included in the artificial neural network model of this study. Consequently, the deaths of the patients suffering from early breast cancer within 5 years cannot be accurately predicted. Therefore, we recommend the inclusion of cytopathological variables such as the determination of the degree of differentiation (i.e., Grades I, II, and III) based on the pathological sectioning of the malignant cells, familial medical history, hormone receptors ER, PR, and HER2 in future studies. Correcting the prediction model with malignant factors that can cause death among patients with early breast cancer within 5 years should lower the incidence of false positive cases and improve the overall prediction of survival rates [2,4,6].

REFERENCES

[1] Department of Health, Executive Yuan. 2011 statistics of leading cause of death in Taiwan, 2012.

[2] P. J. Roohan, , N. A. Bickell, M.S. Baptiste, G.D. Therriault, E.P. Ferrara, A.L.Siu, Hospital volume differences and five-year survival from breast cancer. Am J Public Health, vol. 88, pp. 454-7, 1998.

[3] M. Lundin, J. Lundin, H.B. Burke, S. Toikkanen, L. Pylkkänen, H. Joensuu, Artificial neural networks applied to survival predication in breast cancer. Oncology, vol. 57, pp. 281-6, 1999.

[4] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med, vol. 34, pp. 113-27, 2005.

[5] Y. F. Lu, Predicting Breast Cancer Patients' Survivability: The comparison of Using Three Data Mining Methods- Artificial Neural Network, Logistic Regression and Decision Tree. A master thesis of Division of Epidemiology, School of Public Health, National Defense Medical Center, 2006.

[6] M. J. Van de Vijver, Y. D. He, L.J. van't Veer et al., A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med, vol. 347, pp. 1999-2009, 2002.