Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction

Harry B. Burke, M.D., Ph.D.[1]
Philip H. Goodman, M.D., M.S.[2]
David B. Rosen, Ph.D.[1]
Donald E. Henson, M.D.[3]
John N. Weinstein, M.D., Ph.D.[4]
Frank E. Harrell, Jr., Ph.D.[5]
Jeffrey R. Marks, Ph.D.[6]
David P. Winchester, M.D.[7]
David G. Bostwick, M.D.[8]

[1] Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, New York.

[2] Department of Medicine, University of Nevada School of Medicine, Reno, Nevada.

[3] Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, Maryland.

[4] Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland.

[5] Department of Health Evaluation Sciences, University of Virginia School of Medicine, Charlottesville, Virginia.

[6] Department of Surgery, ɪke Universitv, Durham, North Carolina.

[7] Department of Surgery, Evanston Hospital, Evanston, Illinois; Commission on Cancer, American College of Surgeons, Chicago, Illinois.

[8] Department of Pathology, Mayo Clinic and Mayo Foundation, Rochester, Minnesota.

BACKGROUND. The TNM staging system originated as a response to the need for an accurate, consistent, universal cancer outcome prediction system. Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified and new methods for integrating prognostic factors have been developed. This study compares the prediction accuracy of the TNM staging system with that of artificial neural network statistical models.

METHODS. For 5-year survival of patients with breast or colorectal carcinoma, the authors compared the TNM staging system's predictive accuracy with that of artificial neural networks (ANN). The area under the receiver operating characteristic curve, as applied to an independent validation data set, was the measure of accuracy.

RESULTS. For the American College of Surgeons' Patient Care Evaluation (PCE) data set, using only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis), the artificial neural network's predictions of the 5-year survival of patients with breast carcinoma were significantly more accurate than those of the TNM staging system (TNM, 0.720; ANN, 0.770; $P <$ 0.001). For the National Cancer Institute's Surveillance, Epidemiology, and End Results breast carcinoma data set, using only the TNM variables, the artificial neural network's predictions of 10-year survival were significantly more accurate than those of the TNM staging system (TNM, 0.692; ANN, 0.730; $P < 0.01$). For the PCE colorectal data set, using only the TNM variables, the artificial neural network's predictions of the 5-year survival of patients with colorectal carcinoma were significantly more accurate than those of the TNM staging system (TNM, 0.737; ANN, 0.815; $P < 0.001$). Adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network's predictions of breast carcinoma survival (0.784) and colorectal carcinoma survival (0.869).

CONCLUSIONS. Artificial neural networks are significantly more accurate than the TNM staging system when both use the TNM prognostic factors alone. New prognostic factors can be added to artificial neural networks to increase prognostic accuracy further. These results are robust across different data sets and cancer sites. *Cancer* 1997; 79:857–62. © *1997 American Cancer Society.*

KEYWORDS: TNM staging system, artificial neural networks, prognostic factors, breast carcinoma, colorectal carcinoma, survival, outcomes, decision-making, clinical trials, quality assurance.

ʾresented at the annual meeting of the Ameriːan Joint Committee on Cancer, Scottsdale, Ariːona, January 14, 1995.

Supported in part by research grants from the American Cancer Society (CCG-274), the National Cancer Institute (CA 11606-17), the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383), the Agency for Health Care Policy and Research (HS 06830), and the American Joint Committee on Cancer.

The authors thank John H. Hellier, M.S. ˈor his assistance on this project.

Address for reprints: Harry B. Burke, M.D., Ph.D., Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, NY 10595.

Received May 3, 1996; revision received Octc ber 4, 1996; accepted October 15, 1996.

The TNM staging system originated as a response to the need for an accurate, consistent, universal cancer outcome prediction system.[1] Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified[2,3] and new methods for integrating prognostic factors have been developed.[3] These methods may be capable of (1) providing more accurate predictions than the TNM staging system, using the TNM variables alone (primary tumor size, regional lymph node involvement, and distant metastasis), and (2) further increasing prognostic accuracy by integrating new prognostic factors with the TNM variables. This study compares the cancer specific 5-year survival prediction accuracy for breast and colorectal carcinoma of the TNM staging system with that of artificial neural network statistical models.

METHODS

Data

We used the Commission on Cancer's breast and colorectal carcinoma Patient Care Evaluation (PCE) data sets and the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) breast carcinoma data set.

In October 1992, the American College of Surgeons (ACS) requested cancer information from ACS-accredited hospital tumor registries in the United States. Specifically, they requested the first 25 cases of first-diagnosis breast and colorectal carcinoma seen at each institution in 1983, as well as follow-up information, including deaths, through the date of the request. Variables from this data set used in the breast carcinoma analysis were age, race, payment method, menopausal status, family history, previous biopsy, other cancer, other breast carcinoma, nipple discharge, mammogram, where in the breast the carcinoma occurred, necrosis, histologic grade, estrogen receptor status, progesterone receptor status, number of lymph nodes positive, number of lymph nodes examined, presence or absence of distant metastasis, tumor size, tumor type (in situ, extension to chest wall, or inflammatory), treatment (surgery, chemotherapy, or radiation therapy), and patient outcome (alive or dead). All variables were binary except age, tumor size, number of positive lymph nodes, and number of lymph nodes examined. The PCE data set contained up to 8 years of follow-up information. The analysis end point was breast carcinoma specific 5-year survival. Cases with missing data and those censored before 5 years were excluded. The data set was randomly divided into a training set of 5169 cases, including training and stop-training subsets, and a validation set of 3102 cases.

Variables from the PCE data base used in the colorectal carcinoma analysis were age, race, gender, signs and symptoms (changes in bowel habits, obstruction, jaundice, malaise, occult blood, abdominal pain, pelvic pain, rectal bleeding, or others), diagnostic and extent-of-disease tests (endoscopy, radiography, barium enema, computed tomography scan, biopsy, carcinoembryonic antigen, X-ray, colonoscopy, flexible sigmoidoscopy, intravenous pyelography, liver function tests, biopsy, or other tests), primary site of tumor, level of tumor, histology, grade, number of lymph nodes examined, number of lymph nodes positive, distant metastases, and patient outcome (alive or dead). The end point was 5-year colorectal carcinoma specific survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 5007 training cases, including training and stop-training subsets, and a validation set of 3005 cases.

The National Cancer Institute's SEER breast carcinoma data set, for new cases collected from 1977–1982, with 10-year follow-up, was also analyzed. The extent-of-disease variables for the SEER data set were comparable to, but not always identical with, the TNM variables. The end point was breast carcinoma specific 10-year survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 3788 training cases, including training and stop-training subsets, and a validation set of 2999 cases.

Models

The TNM staging system used in this analysis was the pathologic system based on the American Joint Committee on Cancer's *Manual for Staging of Cancer.*[1] The TNM staging system's predicted survival for a patient in a particular stage is the average survival of patients in that stage.

In medical research, the most commonly used artificial neural networks (ANN) are multilayer perceptrons that use backpropagation training (Figure 1). Backpropagation consists of fitting the parameters (weights) of the model by a criterion function, usually squared error or maximum likelihood, using a gradient optimization method. In backpropagation artificial neural networks, the error (the difference between the predicted outcome and the true outcome) is propagated back from the output to the connection weights in order to adjust the weights in the direction of minimum error. (For a more detailed description of artificial neural networks, see Burke[4] and Cross.[5]) The artificial neural network employed in this research was composed of three interconnected layers of nodes: an input layer, with each input node corresponding to a patient variable; a hidden layer; and an output layer. All nodes after the input layer sum the inputs to them and use a transfer function (also known as an activa-

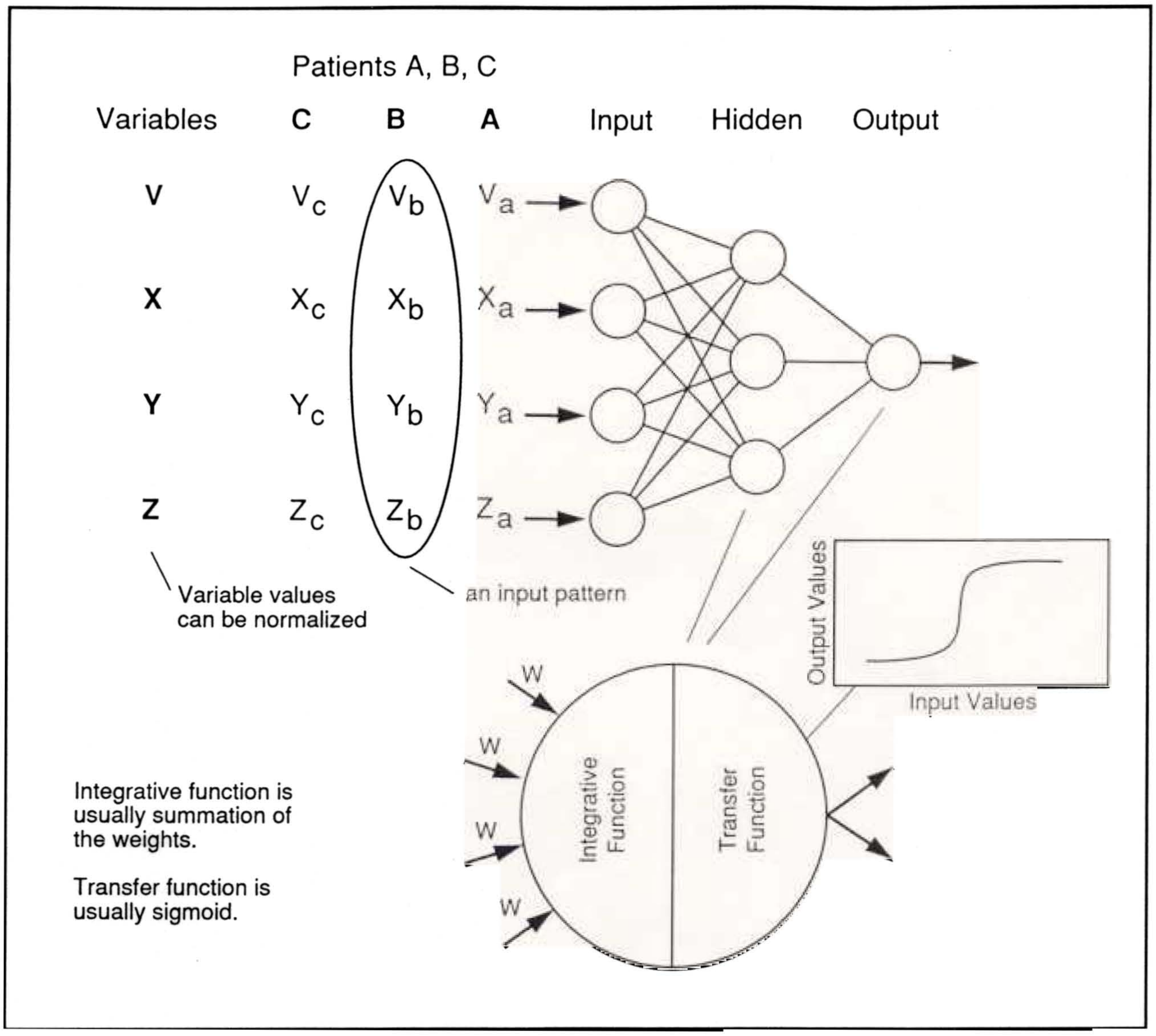

FIGURE 1. Patient A's variable values (Va–Za) are entered into the artificial neural network, followed by patient B, etc. Each variable's input value is multiplied by the weight between the input node for that variable and each hidden layer node it is connected to. All the weighted values going to a hidden layer node are summed at the hidden layer node and go through a sigmoid function before being transferred to the output node. All the weighted values coming into the output node are again summed and put through a sigmoid function. For each patient, the output is a probability from 0–1.0. In training the artificial neural network, the output of each patient is compared with each patient's true outcome. The weights are adjusted so that the next time the patient is presented to the network, the network output is closer to the true outcome.

tion function) to send the information to the adjacent layer nodes. The transfer function is usually a sigmoid function, e.g., the logit. The connections between the nodes have adjustable weights that specify the extent to which the output of one node will be reflected in the activity of the adjacent layer nodes. These weights, along with the connections among the nodes, determine the output of the network.

The mathematical representation of an artificial neural network shown here is equivalent to the graphi model in Figure 1:

$$h_j = f(w^h_{j1}x_1 + w^h_{j2}x_2 + \qquad + w^h_{jn}x_n) \tag{1}$$

$$o_i = g(w^o_1h_1 + w^o_2h_2 + \qquad + w^o_nh_n) \tag{2}$$

where "h_j," in Equation 1 is the output of each of the hidden nodes j, f is a nonlinear transfer function, w^h is the weight from predictor i to hidden node j, and

x_i is an input variable. In Equation 2, o_j is the prediction of the network, g is a nonlinear transfer function, w^o is the weight to the output node, and h is the hidden node output. It should be noted that Equation 2, without the input from Equation 1, is equivalent to logistic regression, where g is the logistic function, w is the beta coefficient, and h is the x covariate.

Specifically, our artificial neural network (NevProp software implementation) used backpropagation training, the maximum likelihood criterion function, and a gradient descent optimization method. The number of input nodes correspond to the number of input variables, the number of hidden layer nodes ranged from three to five, and there was one output mode. Significant differences in the receiver operating characteristic areas between the TNM staging system and the artificial neural network were tested according to the method of Hanley and McNeil.[6] The training data set was divided into training and stop-training subsets. (Training was stopped when accuracy started to decline on the stop-training data subset.) All analyses employed the same training and validation data sets, and all results were based on the one-time use of the validation data sets.

Accuracy

There are three components to predictive accuracy: the amount and quality of the data, the predictive power of the prognostic factors, and the prognostic method's ability to capture the power of the prognostic factors. This study focused on the third component.

The measure of comparative accuracy is the trapezoidal approximation to the area under the receiver operating characteristic curve.[7] The area under this curve is a nonparametric measure of discrimination. While squared error summarizes how close each patient's prediction is to the true outcome, the receiver operating characteristic area measures the relative goodness of the set of predictions as a whole by comparing the predicted probability of each patient with that of all pairs of patients. This area is calculated using the predictive scores of each algorithm in order to compare their average accuracy in predicting outcome. The receiver operating characteristic area is independent of both the prior probability of each outcome and the threshold cutoff for categorization, and its computation requires only that the algorithm produce an ordinally-scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the algorithm will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from 0 to 1. When the prognostic score is unrelated to survival, the score is 0.5, indicating chance accuracy. The farther the score is from 0.5, the better, on average, the prediction model is at predicting which of the two patients will be alive.

TABLE 1
Comparison of the TNM Staging System with the Artificial Neural Network

Data sets	TNM staging system	Artificial neural network
PCE breast CA, TNM variables alone	0.720	0.770[a]
PCE breast CA, TNM and added variables	0.720	0.784[a]
SEER breast CA, TNM variables alone	0.692	0.730[b]
PCE colorectal CA, TNM variables alone	0.737	0.815[a]
PCE colorectal CA, TNM and added variables	0.737	0.869[a]

PCE: Patient Care Evaluation (Commission on Cancer); SEER: Surveillance, Epidemiology, and End Results (National Cancer Institute).

[a] $P < 0.001$.

[b] $P < 0.01$.

RESULTS

A comparison of the accuracy of the TNM staging system and the artificial neural network is shown in Table 1. For the PCE breast carcinoma data set, using only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis), the artificial neural network's predictions of breast carcinoma specific 5-year survival were significantly more accurate than those of the TNM staging system (TNM 0.720; vs. ANN, 0.770, $P < 0.001$). Since the TNM staging system is, by definition, limited to the TNM variables, additional variables do not improve the TNM staging system's predictive accuracy. However, adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network (to 0.784).

We were able to test whether the artificial neural network's significant improvement in predictive accuracy was generalizable across data sets. For the National Cancer Institute's 1977–1982 SEER breast carcinoma data set, using only the TNM variables, the artificial neural network's predictions of 10-year survival were significantly more accurate than those of the TNM staging system (TNM 0.692 vs. ANN 0.730, $P < 0.01$).

We were able to test whether the artificial neural network's significant improvement in predictive accuracy was generalizable across cancer sites. For the PCE colorectal data set, using only the TNM variables, the artificial neural network's predictions of 5-year colorectal carcinoma specific survival were significantly more accurate than those of the TNM staging system (TNM 0.737 vs. ANN 0.815, $P < 0.001$). Adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network (0.869).

To clarify the clinical importance of the observed increases in accuracy, we changed the area under the curve (A_z) scale to a −1 to +1 scale, i.e., $[2(A_z - 0.5)]$. On this scale, 0 was chance and 1.0 was perfect prediction. By this measure, the TNM staging system's accuracy was 44% greater than chance for breast carcinoma specific 5-year survival predictions. Placing the TNM variables in the artificial neural network increased predictive accuracy to 54%, and adding variables that individually had little prognostic value to the artificial neural network further increased prognostic accuracy to 57% greater than chance prediction. Corresponding increases in predictive accuracy specific to colorectal carcinoma were as follows: 47% for the TNM staging system increased to 63% when the TNM variables were placed in the artificial neural network, and that increased to 74% when several commonly collected variables were added to the artificial neural network.

DISCUSSION

The TNM staging system is only moderately accurate in its breast and colorectal carcinoma specific 5-year survival predictions. The significant superiority in predictive accuracy that the artificial neural network showed when compared with the TNM staging system across data sets and cancer sites suggests that it is able to improve our ability to predict the survival of cancer patients. In addition, artificial neural networks can be expanded to include any number of prognostic factors. They can accommodate continuous variables and they can provide presurgery and postsurgery treatment predictions.

Artificial neural networks are a class of nonlinear regression and discrimination statistical methods. They are of proven value in many areas of medicine.[8–19] They do not require a priori information regarding the phenomenon, and they make no distributional assumptions. When the appropriate method is used to avoid overfitting (i.e., loss of generalization by fitting the patterns to the test data too precisely), artificial neural networks are usually at least as accurate as classical statistical models, and, depending on the complexity of the phenomena, they can be much more accurate. In predicting 5-year breast carcinoma specific survival, they have been shown to be more accurate than logistic regression, classification and regression trees (CART; pruned or shrunk), and principal components analysis.[20]

The improvement in prognostic ability made possible by artificial neural networks may be clinically important for therapy, clinical trials, patient information, and quality assurance. In decision-making regarding therapy, it may allow the efficient separation of patients with a poor prognosis (who require therapy) from patients with an excellent prognosis (who require little or no therapy), and it may predict who will respond to a particular therapy. In clinical trials, it may decrease interpatient variability. This would allow for the creation of more homogenous patient populations for clinical trials, resulting in smaller clinical trial patient populations, less expensive trials, and the ability to detect treatment effects that would be undetectable in more heterogeneous study populations. With regard to patient information, it may give patients a clearer understanding of the time course of their disease. Finally, for assessment and quality assurance, it may provide a better severity of illness adiustment.

REFERENCES

1. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors. American Joint Committee on Cancer. Manual for staging of cancer. 4th edition. Philadelphia: JB Lippincott, 1992.
2. Burke HB, Hutter RVP, Henson DE. Breast carcinoma. In: P Hermanek, MK Gospadoriwicz, DE Henson, RVP Hutter, LH Sobin, editors. UICC prognostic factors in cancer. Berlin: Springer-Verlag, 1995: 165–76.
3. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* 1993;72:3131–5.
4. Burke HB. Artificial neural networks for cancer research: outcome prediction. *Semin Surg Oncol* 1994;10:1–7.
5. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet* 1995;346:1075–9.
6. Hanley JA, McNeil BJ. The meaning of the use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic. *J Math Psy* 1975;12;387–415.
8. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995;346:1135–8.
9. Dybowski R, Gant V. Artificial neural networks in pathology and medical laboratories. *Lancet* 1995;346:1203–7.
10. Westenskow DR, Orr JA, Simon FH. Intelligent alarms reduce anesthesiologist's response time to critical faults. *Anesthesiology* 1992;77:1074–9.
11. Tourassi GD, Floyd CE, Sostman HD, Coleman RE. Acute pulmonary embolism: artificial neural network approach for diagnosis. *Radiology* 1993;189:555–8.
12. Leong PH, Jabri MA. MATIC: an intracardiac tachycardia classification system. *Pacing Clin Electrophysiol* 1992; 15:1317–31.
13. Gabor AJ, Seyal M. Automated interictal EEG spike detection using artificial neural networks. *Electroencephalogr Clin Neurophysiol* 1992;83:271–80.
14. Goldberg V, Manduca A, Ewert DL. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Med Phys* 1992; 19:1275–81.
15. O'Leary TJ, Mikel UV, Becker RL. Computer-assisted image interpretation: use of a neural network to differentiate tubular carcinoma from sclerosing adenosis. *Mod Pathol* 1992;5:402–5.
16. Dawson AE, Austin RE, Weinberg DS. Nuclear grading of breast carcinoma by image analysis. *J Clin Pathol* 1991; 95(Suppl):S29–S37.

17. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology* 1993;187:81–7.
18. Astin ML, Wilding P. Application of neural networks to the interpretation of laboratory data in cancer diagnosis. *Clin Chem* 1992;38:34–8.
19. von Osdol W, Myers TG, Paull KD, Kohn KW, Weinstein JN. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J Natl Cancer Inst* 1994;86:1853–9.
20. Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In: Tesauro G, Touretzky DS, Leen TK, editors. Advances in neural information processing systems 7. Cambridge, MA: MIT Press, 1995: 1063–7.