Conference on Prognostic Factors and Staging in Cancer Management: Contributions of Artificial Neural Networks and Other Statistical Methods

Supplement to Cancer

Prediction of Individual Patient Outcome in Cancer

Comparison of Artificial Neural Networks and Kaplan–Meier Methods

David G. Bostwick, M.D., M.B.A.[1,2,3,4,5]
Harry B. Burke, M.D., Ph.D.[6,7]

[1] Bostwick Laboratories, Richmond, Virginia.

[2] Urology of Virginia, Norfolk, Virginia.

[3] Virginia Urology Center, Richmond, Virginia.

[4] Urologic Oncology Division, National Cancer Institute, Bethesda, Maryland.

[5] University of Virginia, Charlottesville, Virginia.

[6] Department of Medicine, New York Medical College, Valhalla, New York.

[7] Department of Oncology, New York Medical College, Valhalla, New York.

BACKGROUND. There is a great need for accurate treatment and outcome prediction in cancer. Two methods for prediction, artificial neural networks and Kaplan–Meier plots, have not, to the authors' knowledge, been compared previously.

METHODS. This review compares the advantages and disadvantages of the use of artificial neural networks and Kaplan–Meier curves for treatment and outcome prediction in cancer.

RESULTS. Artificial neural networks are useful for prediction of outcome for individual patients with cancer because they are as accurate as the best traditional statistical methods, are able to capture complex phenomena without a priori knowledge, and can be reduced to a simpler model if the phenomena are not complex. Kaplan–Meier plots are of limited accuracy for prediction because they require partitioning of variables, require cutting continuous variables into discrete pieces, and can only handle one or two variables effectively.

CONCLUSIONS. Artificial neural networks are an efficient statistical method for outcome prediction in cancer that utilizes all available powerful prognostic factors and maximizes predictive accuracy. Use of Kaplan–Meier plots for predictions is discouraged because of serious technical limitations and low accuracy. ***Cancer* 2001;91:1643–6.** *© 2001 American Cancer Society.*

KEYWORDS: outcome, survival, artificial neural network, Kaplan–Meier, life table, prediction, prognosis.

Predictive factors, when placed in a statistical model, can stratify patients into groups for treatment. There is an acute need for accurate predictive models for prostate carcinoma and all cancers.

Several statistical methods can be used for outcome prediction including descriptive methods (e.g., TNM stages, life table, and Kaplan–Meier approaches) and inferential methods (e.g., logistic regression, proportional hazards, and artificial neural networks). Neural network models provide greater accuracy for combinations of predictive factors than traditional statistical methods of analysis such as the TNM staging system.[1] In this report, we compare artificial neural networks to the recent use of published Kaplan–Meier plots for cancer outcome prediction.

ARTIFICIAL NEURAL NETWORK OUTCOME PREDICTION

Artificial neural networks are a nonlinear regression statistical method that is of proven value in cancer.[1] Artificial neural networks are universal approximators; with sufficient hidden units they can approximate any continuous function to the extent that it is repre-

Presented at the Conference on Prognostic Factors and Staging in Cancer Management: Contributions of Artificial Neural Networks and Other Statistical Methods, Arlington, Virginia, September 27–28, 1999.

Address for reprints: David G. Bostwick, M.D., M.B.A., Bostwick Laboratories, 2807 North Parham Road, Richmond, VA 23294; Fax (804) 288-6568; E-mail: bostwick@bostwicklaboratories.com

Received December 18, 2000; accepted January 26, 2001.

TABLE 1
Comparison of Artificial Neural Networks and Kaplan-Meier Plots for Outcome Prediction

Artificial neural networks
Very efficient statistical method
Uses all the powerful prognostic factors
Maximizes predictive accuracy
Kaplan-Meier plots
Requires mutually exclusive and exhaustive partitioning of the prognostic factors
Limited by lack of patients in most bins
Variables are unequal in predictive accuracy
Reduces accuracy by cutting continuous variables into discrete pieces ("cut points")
Usually accommodates only one variable, or, at most, two or three variables

sented in the data.[2,3] In medical research, the most commonly used artificial neural networks use feed-forward learning, backpropagation of the error, and sigmoid transfer functions. An artificial neural network is composed of three layers of nodes: an input layer with each node in the layer representing a variable (a prognostic factor), a hidden layer that combines those values in a nonlinear manner, and an output layer that provides the probability of an event. The connections between the nodes across layers have adjustable weights that specify the extent to which the output of one node will be reflected in the activity of the subsequent nodes. Backpropagation consists of fitting the parameters (weights) of the model by some criterion function, usually square error or maximum likelihood, using a gradient optimization method.

Artificial neural networks are recommended for cancer prediction because 1) they are as accurate as the best traditional statistical methods, 2) they are able to capture complex phenomena (e.g., nonmonotonic functions and complex interactions) without a priori knowledge, and 3) if the phenomena is not complex, they can be reduced to a simpler model (Table 1).[4]

There are several possible objections to artificial neural networks. First, they may be overparamaterized because they can have many weights. However, overfitting can be prevented by keeping the weights small, thereby reducing the effective number of degrees of freedom. This can be accomplished by penalizing large weights or stopping the iterative fitting algorithm before the weights have grown large. These methods frequently provide better predictive accuracy than if a smaller model was fitted without restriction. When a method is used that reduces the weights that are not being increased by the input variables, the weights in the hidden layer shrink and, when there are only linear relations present, as the hidden layer weights approach zero, the neural network approximates a generalized linear model.

A second potential objection is that artificial neural networks are less "transparent" than traditional statistical methods (i.e., the importance of the variables is less obvious than with traditional models). However, this view of transparency misunderstands the situation. Artificial neural networks are as complex as is necessary to capture the phenomenon. Generally, if the phenomenon is complex, the model must be complex. If the phenomenon is simple enough to be captured by simple traditional models, then artificial neural networks can be reduced to a simple model, and the importance of the variables can be easily seen. For example, if the phenomenon is linear, then a two layer (no hidden layer) artificial neural network with linear transfer functions is mathematically identical to linear regression, and the weights of the artificial neural network are identical to the beta coefficients of the linear regression model. Therefore, model transparency is best understood as a function of complexity and accuracy rather than an intrinsic property of a model. For simple phenomena, a properly chosen simple model is easily interpreted. For complex phenomena (e.g., one that requires complex interactions), a complex model is required to accurately represent the phenomena. Increases in model complexity reduce the transparency of both traditional statistical models and artificial neural network statistical models.[4]

KAPLAN–MEIER OUTCOME PREDICTION

Can individual patient predictions be made by copying life Kaplan–Meier plots from published studies without access to the original data? This approach to data management currently is being used, but it may be harmful or, at the very least, misleading for individual patients. It is wrong to suggest that predictions derived from Kaplan–Meier methods are "individualized" or "personalized"; instead, they are group predictions. In one recent application, the proponents selected published articles that contain Kaplan–Meier curves with one or two predictive factors and then applied a ruler and went year by year from the x axis to the points on the Kaplan–Meier curve, reading the numbers off the y axis, and putting these numbers into a table of survival probabilities. When a patient's predictive factors are entered into the system, the tables are searched, and the one(s) that contains at least one of the patient's factors are found, presumably providing the patient with the annual percentage probability for that outcome over a given time interval. Use of such methods is discouraged by the Amer-

ican Joint Committee on Cancer (G.P. Murphy, personal communication).

How useful are Kaplan–Meier plots and life table plots for making clinical outcome predictions? These methods provide descriptive estimates of the survival of a population of patients over time. The important difference between the life table method and the Kaplan–Meier method is that the life table method provides outcome estimates (e.g., survival) at uniform time intervals (e.g., at each year), whereas the Kaplan–Meier method provides outcome estimates whenever an event occurs (e.g., a death). These methods are of limited utility for outcome prediction for several reasons (Table 1).

First, both methods are "bin" methods. Bin methods require mutually exclusive and exhaustive partitioning of the variables so that every patient is in one and only one bin (cases in which a bin is a unique set of patient variable values, e.g., tumor size, lymph node number, and presence or absence of distant metastasis). Each bin contains a subset of the patients from the overall population. Each bin's prediction is the outcome of the patients in that bin over time. Thus, if 50% of the patients in a bin are alive at 5 years, then that bin predicts a 50% 5-year survival.

Second, the bins used by life table plots and Kaplan–Meier plots are "lumpy." This means that there is a high and disproportionate number of patients in a few bins and few or no patients in most bins. There is no connection between bins, so the bins with patients cannot "help" the bins with few or no patients. Thus, for many patients, accurate prediction is not possible.

Third, these methods often compress data and are limited by the number of variables that can be accommodated. Life table and Kaplan–Meier estimates usually are used in medicine to determine if a variable, when cut into two or three pieces (cut points), separates a patient population into two or three groups. All too often, these cut points are chosen arbitrarily or at a convenient threshold, notwithstanding the statistical analyses needed to optimize the cut points. If the separation is significant, that suggests that the variable may play a role in the disease process. With this approach, any increase in the number of variables or categories within each variable must be matched by an *exponential increase* in the number of patients to maintain predictive accuracy. Thus, with as few as two or three variables, this approach requires high numbers of patients—adding variables becomes impossible. Furthermore, two or three variables are almost never adequate for accurate prediction in cancer. Finally, this means that all the variables must be discrete. Continuous variables such as tumor size, number of positive lymph nodes, and even age must be cut into a few pieces. There is a loss of predictive accuracy when continuous variables are cut into pieces (a truism of statistics). Predictions using this approach will almost always be based on one, two, or at most three discrete variables. The low number of variables and the discrete nature of the variables will result in predictions that are not particularly accurate, and therefore not clinically useful. Thus, predictions should not be made from copies of plots that were published.

Fortunately, life table and Kaplan–Meier estimates are rarely used for patient outcome predictions; when used for this purpose, they must be specially constructed using the raw data from a properly defined and collected patient population. However, most contemporary applications that use life tables and Kaplan–Meier estimates are not undertaking this critical step.

There are several dangers associated with use of the published literature without access to the raw data. One problem is the comparability of outcomes within a treatment. Another problem is the comparability of the outcomes across treatments. It is worthwhile to consider examining within-treatment comparability of life table and Kaplan–Meier plots of survival for a patient considering radical prostatectomy for prostate carcinoma. Medline searches reveal more than 1400 studies published on survival after radical prostatectomy. Kaplan–Meier proponents usually cite only a few of these studies; most studies would have to be rejected because they do not contain Kaplan–Meier plots or do not contain the variables of interest. These are very different studies, and that is reflected in their very different results. They test different hypotheses, they differ in size, they contain different patient populations with different demographics and different severities of illness, their patients were collected at different times and in different ways, they have different rates of lost to follow-up (patients that were censored), their patient variables are defined differently, their outcomes are defined differently, and they use different combinations of patient variables. Furthermore, many studies find different and conflicting results. Furthermore, in most the data have not been validated with independent data sets.

The problem becomes worse when one attempts to compare two outcomes such as recurrence and survival for the same therapy. Recurrence and survival predictions are based on different studies, so there is no necessary relation between these outcomes for the same patient.

Other severe problems are encountered when attempting to compare treatments. Comparative treatment benefit is what many patients want to know. The

patient populations for different therapies are usually very different. For example, it is well known that patients with operable prostate carcinoma may receive either surgery or radiation and that patients with inoperable prostate carcinoma may receive radiation therapy. How can a radiation therapy study be reliably compared with a radical prostatectomy study using the life table or Kaplan–Meier approach? It cannot. One must conclude that the presentation of life table or Kaplan–Meier plots from different studies for different therapies is extremely misleading and cannot be justified.

Survival predictions based on studies that contained only clinical stage, studies that contained only Gleason score, or studies that contained only prostate specific antigen will have a poor predictive accuracy. What is required is a model that can combine stage, Gleason score, and prostate specific antigen. Because each strata of each variable must be represented by an outcome curve, the proliferation of curves quickly becomes untenable and the results useless.

Presenting copied life table and Kaplan–Meier plots from a few heterogeneous published articles does not provide patients with reliable, accurate, and useful clinical predictions. It is better to provide accurate inferential model-based estimates using artificial neural networks.

REFERENCES

1. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr., et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997;79: 857–62.
2. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw* 1989;2: 359–66.
3. Leshno M, Lin VY, Pinkis A, Schocken S. Multilayer feedforward networks with nonpolynomial activation function can approximate any function. *Neural Netw* 1993;6:861–7.
4. Burke HB. Statistical analysis of complex systems in biomedicine. In: Fisher D, Lenz H, editors. Learning from data: artificial intelligence and statistics V. New York: Springer-Verlag, 1996:251–8.