Evaluating prognostic factors

HARRY B. BURKE, M.D.[1],
DONALD E. HENSON, M.D.[2]

Bioinformatics and Health Services Research, Department of Medicine[1], New York Medical College, Valhalla New York, Cancer Biomarkers Research Group, Division of Cancer Prevention[2], National Cancer Institute, Bethesda, Maryland

ABSTRACT Prognostic factors are necessary and sufficient for assessing the natural history of cancer, selecting the optimal therapy and evaluating the effectiveness of treatment. Because prognostic factors are predictive to the extent that they participate in the disease process, anything that participates in the disease process is a potential prognostic factor. Any investigation of the disease process, therefore, can result in the identification of new prognostic factors. As researchers move down explanatory levels of analysis, and especially when they explore the molecular genetic level, they increase explanatory complexity. One result of this increase in complexity is the proliferation of prognostic factors. In addition, methodologic and technical issues arise related to the identification, replication and validation of molecular genetic factors. The combination of the proliferation of putative factors and the lack of replication and validation of findings has resulted in confusion in the prognostic factor domain.

In this paper we explore some of the reasons for the ambiguity surrounding the non-extent of disease putative prognostic factors and how these ambiguities can be resolved. Specifically, we: 1. define and describe prognostic factors, as a type of predictive factor, 2. explain why, and under what conditions, combining factors may increase predictive accuracy, and 3. describe the advantages and disadvantages of commonly used statistical methods for combining predictive factors, and 4. recommend an approach to the reporting of prognostic factor research results.

Key words ovary, prognostic factor, cancer

Address correspondence to:

Harry B. Burke, M.D.
Bioinformatics and Health Services Research
Department of Medicine, New York Medical College
Valhalla NY 10595, USA
Phone (1 914) 594 4920 Fax (1 914) 594 4923
E-mail harry_burke@nymc.edu.

INTRODUCTION The clinical prediction of patient outcome is based on the integration of one or more prognostic factors in a descriptive or inferential statistical model. Prognostic factors are necessary and sufficient for assessing the natural history of cancer, selecting the optimal therapy and evaluating the effectiveness of treatment (1). Because prognostic factors are predictive to the extent that they participate in the disease process, anything that participates in the disease process is a potential prognostic factor. Any investigation of the disease process, therefore, can result in the identification of new prognostic factors. As researchers move down explanatory levels of analysis, and especially when they explore the molecular genetic level, they increase explanatory complexity. One result of this increase in complexity is the proliferation of prognostic factors. In addition, methodologic and technical issues arise related to the identification, replication and validation of molecular genetic factors. The combination of the proliferation of putative factors and the lack of replication and validation of findings has resulted in confusion in the prognostic factor domain (2-3).

BACKGROUND The most commonly used ovarian cancer prognostic factors have been those that code for the extent of disease at diagnosis, namely, tumor location(s), lymph node involvement and metastasis. These factors are combined into comparable outcome-based groups by the International Federation of Gynecology and Obstetrics' (FIGO) stages and by the International Union Against Cancer's (UICC) TNM categories (4). Performance status is a staging system that depends on functional rather than pathologic variables.

Prognostic factors can be classified by their level of analysis; demographic, anatomic-cellular and molecular-genetic (5). Age is a demographic factor and, like most demographic factors, is not directly related to the disease process. Age is indirectly related to the disease in the sense that it can be a surrogate for time related disease processes and can therefore be weakly prognostic for disease-specific events. Sex and race, to the extent that the disease does not differ in kind by sex or race, are demographic factors. Clearly, if the disease is different in different sexes or races, then two diseases exist.

The majority of the current ovarian cancer prognostic factors exist at the anatomic-cellular level, including the FIGO/TNM factors. Most anatomic-cellular level factors are clinically equivocal for one or more of the following reasons: 1. the factor exhibits high assay variance or high intra or inter-observer variance, 2. the factor possesses low univariate or multivariate predictive power, and 3. there is too much inter-study variation in the factor's assessment, for example, differences in experimental methods, patient populations and outcomes. Ascites and volume of residual disease can be viewed as extent of disease factors that code for advanced disease. Histologic subtype has a subjective component which increases its variance. Because there is no universal grading system there is high inter-observer variance associated with grade. Tumor ploidy, S-phase, DNA index and morphometric measures have been widely investigated, as have various serum factors including CA 125, CA 54/61, and LDH. There is a great deal of inter-study variation is the assessment of these factors.

Recently molecular genetic factors have been investigated including, p53 (6-9), BRCA1 (10), c-erbB-2 (11), bcl-2 (12) and factors that are thought to be related to "drug resistance", including GSTpi expression (13), Lrp (14), and MRP (15). There has been a great deal of controversy regarding the validity of molecular genetic factors and none are universally accepted. An example of why it is difficult to accept these factors is shown for p53 *(Table 1)*. The p53 studies as a group exhibit a great deal of variation in experimental methods, patient populations, associated independent variables, outcomes and other aspects of their design and execution.

In this paper we explore some of the reasons for the ambiguity surrounding the non-extent of disease putative prognostic factors and how these ambiguities can be resolved. Specifically, we: 1. define and describe prognostic factors, as a type of predictive factor, 2. explain why, and under what conditions, combining factors may increase predictive accuracy, and 3. describe the advantages and disadvantages of commonly used statistical methods for combining predictive factors, and 4. recommend an approach to the reporting of prognostic factor research results.

PREDICTIVE AND PROGNOSTIC FACTORS A predictive factor predicts an outcome (risk of disease, existence of disease or prognosis) by virtue of its relationship with the disease process that causes the outcome. Terms such as marker, biomarker, predictor, prognosticator, indicator, surrogate factor and intermediate biomarker have been used to identify variables that are connected to medical outcomes. The meanings of these terms overlap and their undifferentiated use can cause confusion. We suggest that all predictive factors are markers of disease (i.e. they are in some way associated with the disease process), but that not all markers of disease have sufficient predictive power to be called predictive factors. We will use the term factor to identify markers of disease that either are, or have the potential to be, predictive for a given outcome in a specified statistical model.

There are three types of predictive factors; risk, diagnostic, and prognostic. They differ in their outcomes and in their degree of factor-outcome association. "Risk" is an ambiguous term. We will use "risk" to refer to "risk of disease". "Risk" when used in the context of "risk of recurrence" or "risk of death" will be called "probability", as in "probability of recurrence" and "probability of death". A risk factor's main outcome is incidence of disease. The factor, either alone, or in combination with other factors, is much less than 100% predictive of the disease occurring by a specified time in the future. Risk can be viewed as a propensity for the disease. A high-grade squamous intraepithelial lesion (HSIL), for example, is a cytologic risk factor for subsequent cervical cancer. It indicates a greater propensity for cervical cancer than a normal Papanicolaou smear.

A diagnostic factor's main outcome is also incidence of disease. The factor, either alone, or in combination with other factors, is close to 100% predictive of disease. A biopsy that shows invasive cancer is 100% predictive of the disease.

A prognostic factor's main outcome is usually death. A prognostic factor is rarely a strong predictor in isolation from other prognostic factors. Tumor locations(s) and lymph node involvement are prognostic factors in ovarian cancer.

Within each type of predictive factor there are three subtypes: 1. natural-history, 2. response-to- therapy, and 3. post-therapy. Natural-history predictive factors predict the future occurrence (risk), current existence (diagnosis) or course (prognostic) of a disease without a preceding treatment. For risk and prognosis, natural history should the baseline against which all treatments are tested. An example of a natural-history prognostic factor is any extent-of-disease factor such as tumor size. Response-to-therapy predictive factors assume that there are effective therapies and predict whether the patient will respond to a particular intervention (e.g. chemoprevention or chemotherapy). A therapy-specific factor is, as its name implies, specific to a particular treatment and must be assessed in a population that only received that treatment. An example of a therapy-specific prognostic factor is estrogen receptor status in breast cancer. A natural-history predictive factor may also be a response-to-therapy predictive factor if it changes its value in response to a successful treatment. Post-therapy predictive factors require that patients respond to the intervention; they predict failure of the intervention. Recurrence is a post-therapy prognostic factor.

Determining whether a marker is a predictive factor requires that: 1. the variable be measured in a defined population, 2. the population be followed until enough outcomes have occurred (e.g. deaths), and 3. the relationship between the variable and the outcome be determined. If the variable predicts the outcome with "sufficient" accuracy (where sufficient varies with the question being addressed) in a specified model it is called a predictive factor. If the outcome that is predicted to occur always occurs, we say that the predictive factoı and the outcome are 100% linked, i.e. that the factor has a 100% predictive accuracy.

Study	Tissue	Variable	Variable assay method	Variable coding	% of population p53 positive	Population size	Population characteristics
Eltabbakh et al. (6)	Paraffin block	Nuclear protein overexpression	Immunohisto-chemistry	Negative = no staining, three grades of staining: few cells to <25% cell staining, 25 - <75 cell staining, >75% cell staining	48% positive for p53 overexpression	221	Primary ovarian epithelial carcinoma, all stages
Niwa et al. (7)	Paraffin block	p53 gene DNA fragments for allelic losses and mutations	Single-strand conformational polymorphism (PCR-SSCP)	Positive for allelic loss: het herozygous genotype and the signal intensity of a paired allelic fragment is less than 40% of the other paired segments.Positive for mutations if occur on exons 4-8.	Allelic loss: 18 of 26 cases. Mutations: 14 cases.	67	54 primary, 7 metastatic, 6 after chemotherapy
Klemi et al. (8)	Paraffin block	Nuclear protein overexpression and DNA fragments for mutations	Immunohisto-chemistry and Single-strand conformational polymorphism (PCR-SSCP)	Negative = no staining, two grades of cell staining, two grades of cell staining equivocal <20% cell staining, positive >20% cells staining	44% positive, 19% equivocal for p53 overexpression	136	Primary ovarian epithelial carcinoma, all stages
Viale et al. (9)	Paraffin block	Nuclear protein overexpression	Immunohisto-chemistry	Negative = no staining, positive = >10% cell staining	62% positive	112	Primary ovarian epithelial carcinoma, all stages

Table 1. Comparison of p53 studies

The predictive power of a factor depends on its intrinsic and extrinsic power. The intrinsic predictive power of a factor is related to its "connectedness" to the disease process. "Connected" means associated with the disease process (where "process" subsumes trigger, cause, etc.). The less connected the factor is, the less predictive it is. A direct connection means that the factor is an integral part of the disease process itself. An indirect connection means that it is not an integral part of the disease process, but is related to the disease process such as being a byproduct of the disease process. The extrinsic predictive power of the factor depends on the question being asked, i.e. the specific factor-outcome relationship being examined.

For a specific disease process and outcome, the predictive accuracy of a factor depends on: 1. how closely connected the factor is to the disease process (individual factor power) and the orthoginality of the collected factors (degree of predictive overlap), 2. how easy it is to collect and measure the factor, 3. the degree to which the selected statistical method is able to capture individual factor predictive information and to integrate the information from multiple factors.

CRITERIA FOR PREDICTIVE FACTORS Predictive factors should be: 1. accurate, 2. independent, and 3. useful. Accurate means

Event type and rate	Follow-up duration	Other variables in model	Treatments	Findings	Replication of another researcher's results	Results validated (another researcher with another data set)
Survival, median survival was 53.5 months. No event rate was reported	Median 7 years	Surgical stage, residual disease after primary surgery, histology, grade, age, race	Cytoreductive surgery, platinum-based chemotherapy, non-platinum chemotherapy, radiation, intraperitoneal chronic phosphate	Univariate p53 overexpression (none versus all), p = 0.0498. Multivariate p53 overexpression, adjusting for stage and size of residual tumor burden, p = 0.16	No	No
Survival of 31 cases	Not reported	Stage, histologic grade, response to primary therapy	Surgery, cisplatin/doxorubicin/ cyclophosphamide chemotherapy	p53 overexpression not correlated with survival	No	No
Survival, 93 deaths due to ovarian cancer.	Median 122 months	Stage, S-phase, histologic type and grade, DNA ploidy, age at diagnosis	Radical surgery, radiotherapy, chemotherapy	No correlation between p53 immunostaining result and the number of mutations by PCR-SSCP. Positive p53 staining associated with serous histologic type, S-phase, and histologic grade. Univariate and multivariate p53 overexpression associated with survival p = 0.002, and p = 0.008	No	No
			No	No		
	Mean 46 months	bcl-2, MIB1	Surgery, platinum-based combination chemotherapy	Positive p53 staining associated with grade, stage, residual tumor. Overall survival was associated with p53 overexpression, age, grade, state, serous type, and MIB1	No	No

that the factor is, at its minimum accuracy, a powerful predictor for a subset of a clinical population or, at its minimum accuracy, a modest predictor for a large segment of the population. Independent means that the factor retains a significant predictive value when other predictive factors are added to it in a multivariate model. Useful means that the predictive factor is clinically relevant; that it can affect patient management and therefore outcome. Powerful predictors are not always clinically useful because clinical utility also requires an effective therapy, which may not be available. There are non-clinical utilities, one of which can be termed a "social" utility. The social utility of a factor is its ability to provide information to patients regarding their outcome even when the outcome cannot be changed. It is important to note that it is not necessary to understand the function of the factor in order to use it as a predictive factor.

IDENTIFICATION, REPLICATION, AND VALIDATION OF FACTORS

A putative predictive factor must go through three stages of testing before it can be used clinically. The first stage is the discovery and characterization of the factor. The factor must be unambiguously identified and its predictive linkage to a clinical outcome determined (usually using an appropriate univariate statistical method). Important components of the identification stage are the explicit definition of the factor, the detailed description of the method used to detect it, and the inclusion/exclusion criteria and collection methods for the clinical population that is used to assess the factor.

The second stage is replication. Once a factor has been identified it must be replicated using the original assay method and independent researchers. Other assay methods that are commonly used to detect this type of factor should be employed by both the original researcher and by independent researchers. The idea here is that the original finding should be reproducible across assay methods and researchers using the same defined patient population. Failure to replicate previous results will affect the interpretation and use of the prognostic factor. In addition, the accuracy method that is used to compare the factor across assay methods and researchers must be suitable for the comparison of two statistical models .

The third stage is validation. Validation addresses the issue of the predictive power of the factor in other populations. The factor should be assessed on a well defined independently collected patient population (not the same population that was used for identification and replication). The question being addressed is whether the factor retains its predictive power. In order for a factor to be clinically useful it must be assessed by a method that can be performed in many different types and levels of laboratories and it must be powerful enough to overcome intra-observer, inter-observer and inter-institutional variance.

It is a common practice to transform continuous factors into discrete factors. For example, in ovarian cancer the number of lymph nodes involved, which is a continuous factor, is transformed into a binary category in the FIGO/TNM systems. Transforming a continuous factor into a discrete factor reduces the factor's predictive accuracy (1). If a factor must be partitioned, the accuracy of the factor must be assessed on a data set different from the one used to determine the optimal partitioning of the factor (16). Validation in a multivariate model with all the current, related factors is addressed in the next section. Significant findings should be followed-up by a prospective, multi-institution study. There is no need for a control group in such a study.

There are two major validation problems related to prognostic factors. The first is the time from diagnosis to the analysis of outcomes (e.g. mortality). The longer this interval, the longer the prediction time interval. To provide, for example, ten year survival predictions, a patient population must be followed for ten years. The ten-year information is used to assess prognostic factor predictive accuracy and to provide ten-year outcome predictions to future patients. The second is the accrual of a sufficient number of outcomes so that the assessment of the factor is statistically reliable. Reliable means that a similar result would be observed if the analysis were repeated. One solution to these problems is the implementation of a specimen bank (17).

INTRODUCTION TO COMBINING FACTORS It is rarely the case that one factor is sufficiently predictive, i.e. that it is able to predict the outcome of interest with 100% accuracy (until the patient is very near the outcome). The usual strategy when dealing with predictors, especially risk and prognostic factors, is to combine several in a predictive model. The most useful grouping of factors is one in which all the factors are powerful and predictively orthogonal to each other, i.e. they represent independent aspects of the disease process. If they represent aspects of the disease that are not independent then their information will overlap and one will not add predictive power when combined with the other factors. The statistical method employed to combine the factors must be able to capture the complexity of the disease process that is represented by the factors being combined, e.g. non-linearity and interactions.

A predictive model is the result of using a statistical method to relate one or more predictive factors to an outcome. For example, the mathematical formula generated by the logistic regression statistical method relates the predictive factors (input variables), in terms of their ß-coefficients, to a binary disease outcome, e.g. recurrence, death, etc.

It should be noted that the predictive power of a factor is always associated with the statistical method that was used to capture its power and with the other factors included in the model. Because a particular model may not be efficient at capturing the power of the factors, and because not all the relevant factor may be included in the model, any statement of a factor's accuracy must include an explanation of why a specific statistical model was used and why certain factors were included in the model.

METHODS FOR COMBINING FACTORS The primary descriptive methods for combining factors in cancer are: bins, stages and indexes (either as discrete endpoint or as Kaplan-Meier product-limit models) (18). The main inferential methods for combining factors are: decision trees, regression methods including logistic and proportional hazards, and artificial neural networks (19-20).

Bins are the result of the mutually exclusive and exhaustive partitioning of discrete variables. Each combination of variable values is a bin and every patient is placed in the bin corresponding to their variable value combination. An example is the TNM classification of ovarian cancer. Tumor location (T1a, T1b, T1c, T2a, T2b, T2c, T3a, T3b, T3c), regional lymph node involvement (N0, N1) and existence of metastases (M0, M1) produce thirty-six bins.

For discrete variables, if there are enough patients in each bin, it can be shown that the frequency of the outcome in the population within each bin is the best predictor of the true outcome. In other words, no prediction model can be more accurate than the bin model if the variables are discrete and the population very large. Problems with bin models include: 1. continuous variables must be parsed into discrete variables, almost always resulting in a loss of predictive information and therefore a loss of accuracy. 2. As the number of discrete variables increase the number of bins increase exponentially. For example, if we wish to add 3 grades to the TNM of ovarian cancer, then the number of bins will increase to 108. In order to maintain accuracy there must be a corresponding exponential increase in the size of the patient population to fill each bin. 3. The proliferation of bins reduces the ability to understand the phenomena. Since the main reason of creating a bin model is usually for ease of understanding and ease of use, bin models are rarely used in situations where there are more than two or three predictive factors.

A partial solution to some of the problems of a bin model is a stage model. A stage model is combining of bins into super-bins. The justification for grouping is the assumption that the factors selected are indexes of the "stages" of the disease process and that the combined bins represent a real stage in the disease process. For example, in breast cancer, the TNM staging system combines forty TNM classification bins into six super-bins (stages I, IIA, IIB, IIIA IIIB, IV) based on decreasing survival, and these super-bins are termed as the TNM staging system.

A small set of stages have the potential to maintain explanatory simplicity and ease of use. Problems with stage models include: 1. the combining of bins into super-bins/stages reduces predictive accuracy. 2. Stage systems do not overcome the exponential increase in bins and in patients associated with adding a variable to the staging system, they just delay the problem at the cost of predictive accuracy. If the stages are held constant as variables (and their associated bins) are added the staging system, the potential improvement in accuracy associated with the additional bins will be small to nonexistent. But, if the stages are expanded to accommodate additional bins, the system loses its ease of understanding and usefulness. Thus, attempts to improve predictive accuracy by adding variables to a bin/stage model are rarely successful. 3. The problems of parsing continuous variables, with as the resulting loss in predictive accuracy, remains.

Indexes associate numerical scores (usually based on a bounded, linear scale) with bins or groups of bins. The scores are parsed into discrete ranges, and each range is associated with a disease stage (usually a severity of illness system). Indexes offer some flexibility in the grouping of bins, but at the cost of further degradation in predictive accuracy. The simplest example of an index is the Apgar score.

Any bin, group of bins, stages or scores can be contrasted, in terms of outcome, with other bins, group of bins, stages or scores at the end of a single time interval or across a series of event time intervals. (In other words, comparing predictive factors.) Both the single time interval and the event interval approaches usually deal with censoring by dropping censored cases at the time interval in which they are censored. The most common descriptive approach for contrasting predictive factors across a series of event time intervals is the Kaplan-Meier product-limit method (inferential methods that can accommodate continuous variables and that usually require a proportional hazards assumption, will be discussed later with regression methods). A Kaplan-Meier plot should always include confidence intervals around each line. A significant difference in a Kaplan-Meier comparison is usually assessed by a log-rank test (which assumes proportional hazards). It is important to note that there is currently no widely accepted method for comparing the accuracy of two Kaplan-Meier comparisons based on different stratifications of the same variables. The use of the log-rank p-value to select one stratification over another is incorrect because the log-rank test determines whether a factor stratification is likely to have occurred by chance. Extreme stratifications may result in a smaller p-value, but it may also reduce predictive accuracy over the entire population.

Decision trees split predictive factors to maximize predictive power using a loss function such as the log-likelihood and a greedy search algorithm. The most well known decision tree approach is the Classification and Regression Trees (CART) recursive partitioning method (21). Empirically, we have never found decision trees to be the most accurate statistical method, when compared to regression methods. Its problems include the selection of the correct loss function, difficulty dealing with continuous variables, and overfitting when searching for the best predictors especially when there are more than two or three splits.

Univariate regression methods are not appropriate for deciding whether a variable is or is not a predictive factor. These methods should not be used to assert that a factor is predictive because a factor must be assessed in the context of the other known factors. Further, some variables are only predictive when interacting with other factors (for example, some molecular genetic factors).

Logistic regression is the cumulative probability of a binary event occurring by a specific time. It uses a maximum likelihood loss function and a greedy search technique. It is a very efficient method for problems that have a binary outcome (e.g. recurrence, survival). Its limitation is that it must span a single time interval and does not distinguish when the event occurred in the interval. Also, in order to handle censoring one must create a multi-time interval logistic regression model and for each time drop the cases that are censored in that interval.

"Proportional hazards" methods include the Weibull, exponential and Cox. The Cox proportional hazards regression method (22) is the most commonly used. All three methods assume that the hazard of each patient is proportional to the hazards of all the other patients and that the degree of each patient's hazard is related to their relative risk. The Cox model cannot create empirical survival curves. For survival curves, a baseline hazard must be introduced, e.g. Cox-Breslow estimates (23). Some researchers incorrectly believe that the Cox is the only regression method that can deal with censoring. A multi-interval logistic regression can deal with missing data. In cancer, the proportional hazards assumption is often violated. Therefore, anyone using a Cox model must demonstrate that proportional hazards holds for the factors and outcome.

Molecular genetic factors, for example, p53, c-erbB-2 (HER-2/neu), pRB, exhibit the properties of complex systems, they are nonlinear and are inter-actional, i.e. they act non-monotonically and in concert with other molecular genetic factors (24-27). Thus, capturing the factors as part of a complex system is critical to accurate prediction of the behavior of the system. Artificial neural networks are capable of capturing complex systems (28).

The idea that learning can be viewed as the modification of information by repetitively passing it through processing nodes originated in the late 1940's as a way to model the physiology of neuronal processes (29). The operationalization of this idea was called an artificial neural network. Gradually it became apparent that this information theoretic approach to learning was very powerful and very general; it was useful in, and applicable to, many learning situations. Since statistics can be viewed as learning from the data, it is not unexpected that this approach would be mathematically proved and operationalized within the domain of statistics.

Artificial neural networks are universal approximators. It has been shown that any real, continuous function can be approximated to any degree of precision by a three-layer network with x in the input layer (patient variables), a hidden layer with sigmodal transfer functions, and one layer of output units, as long as the hidden layer can be arbitrarily large (30-31).

Artificial neural networks, as a class of nonlinear regression and discrimination statistical methods, are of proven value in many areas of medicine (32-39). They do not require prior information regarding the phenomenon, they make no distributional assumptions, and with the appropriate method 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, can be much more accurate. Artificial neural networks have, for example, been shown to be more accurate than logistic regression, CART (pruned or shrunk) and principal components analysis at predicting five-year breast cancer specific survival (29).

In medical research, the most commonly used artificial neural networks (ANN) are multi-layer perceptrons that use backpropagation training. 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 references 39-40). The usual artificial neural network employed in medical research is 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 activation function) to send the information to the adjacent layer nodes. The transfer function is usually a sigmoid function such as the logistic. 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 output of the network is a probability of the event for each patient.

COMPARING STATISTICAL MODELS: MEASURING ACCURACY In order to assess and compare models, it is necessary to distinguish between significance, accuracy and importance. Significant is the fact, that it is unlikely that either a trained statistical method (i.e. a statistical model) or a predictive factor's predictions are due to chance (e.g. the chi-square test). Significance is not necessarily accuracy. Accuracy is the association between the model's individual patient outcome predictions (the predicted outcome) and the individual outcomes of the test population (the true outcome). The importance of a factor or a model is based on whether the model or factor possesses sufficient accuracy to be useful in answering a particular clinical question. Finally, the assessment of the model's or factor's significance, accuracy and importance must be based on test data set results, not on training data set results.

There are several approaches to assessing the accuracy of a multivariate model and for comparing multivariate models (e.g. Goodman and Kruskall's Gamma, Kendall's Tau). The best method currently in use is the area under the receiver operating characteristic curve. The area under the receiver operating characteristic curve (Az) is the best currently available measure of predictive accuracy (41). It can be used to assess and compare the adequacy of statistical models. Az can be directly calculated by Somer's D (42) or it can be approximated by its trapezoidal area (43). The area under the curve is a non-parametric measure of discrimination. The receiver operating characteristic area is independent of both the prior probability of each outcome and the threshold cutoff for categorization. Its computation requires only that the prediction method produce an ordinal-scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the prediction method will assign a higher mortality score to the patient who died, than to the patient who lived. The receiver operating characteristic area varies from zero to one. When the predictions are 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 method is at predicting which of two patients with different outcomes will be alive. Significant differences in the receiver operating characteristic areas between two models can be tested following *Hanley and McNeil* (44), by calculating their asymptotic variances, or calculating the empirical variance using bootstrap method (45).

UNRESOLVED ISSUES There are several important unresolved issues related to the use of prognostic factors. The first is related to determining the natural history of the disease when effective therapies exist. The problem here is that all therapy-specific factors are compared to the natural history of the disease. But if the natural history is not known, this comparison cannot be made. The second unresolved issue is how to assess new treatments. Specimen banks cannot overcome the need to collect outcome data over long periods of time on large patient populations for new therapies. Perhaps modeling the phenomena and the effect of a new treatment by simulation may be helpful in the future.

REPORTING PREDICTIVE FACTOR RESEARCH RESULTS There is a great deal of variation in reporting of predictive factor results. This variability makes it difficult to understand empirical results and to replicate and validate predictive factor research. A report regarding the discovery of a new predictive factor or the validation of an existing factor should contain the following: 1. Name of disease and where in the disease process the patients are that have been collected (i.e. early detected disease). 2. Name and description of the prognostic factor. 3. Type of prognostic factor (i.e. natural history, therapy-specific, post-therapy). 4. Outcome selected, e.g. five-year breast cancer-specific survival, (it should not be "lifetime" except in special situations). It should be a specific time interval. 5. Time of assay (e.g. at discovery, prior to therapy, after therapy). 6. Specific laboratory method used to assess the factor and why that method was selected (e.g. immunohistochemistry). 7. If the prognostic factor is stratified, the specific range/cut-point/etc. of the factor. If the variable value is based on rater judgment, then Cohen's kappa should be reported. 8. Relevant characteristics of the data set, including: data set size, the number of events, and whether the therapy was randomized. 9. The numerical estimate and confidence interval of the finding. 10. The level of significance, for example, p <0.05 for one test of the data. If multiple tests of the data are performed, an adjustment may be required. 11. The type of multivariate statistical method (e.g. logistic regression, Cox) used and tests for any assumptions (for example, proportional hazards). 12. If all the other relevant prognostic factors were not included in the multivariate model, which were left out. 13. Specific type of therapy, e.g. surgery, chemotherapy and radiation therapy, and how it was admi-nistered. 14. The method of accuracy assessment, i.e. area under the receiver operating characteristic, R2, etc., and why was it used. 15) The accuracy estimates of the multivariate model including either standard errors or confidence intervals for the estimates (e.g. Az = 0.75, CI = 0.50 - 1.0).

In summary, evaluating predictive factors is a complex and difficult process. But it must be done if we are to have true findings that can be used clinically.

REFERENCES

1. Burke HB. Integrating multiple clinical tests to increase predictive accuracy. In: Hanausek M. Walaszek Z. (eds.), Methods in Molecular Biology. Vol. XX: Tumor Marker Protocols.. Tonowa, N.J. Humana Press, 1998a, Chapter 1:3-10.

2. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994; 69:979.

3. Altman DG, Lyman GH. Methodologic challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 1998; 52:289.

4. Fleming ID. Cooper JS, Henson DE, Hutter RVP, et al. AJCC Cancer Staging Manual, 5th ed, Philadelphia, Lippincott-Raven, 1997.

5. Burke HB, Hutter RVP, Henson DE. Breast Carcinoma. In: Hermanek P, Gospadoriwicz MK, Henson DE. Hutter RPV, Sobin LH, (eds.), UICC Prognostic Factors in Cancer. Berlin, Springer-Verlag, 1995:165.

6. Eltabbakh GH, Belinson JL, Kennedy AW, Biscotti CV, Casey G. Tubbs RR, Blumenson LE. p53 overexpression is not an independent prognostic factor for patients with primary ovarian epithelial cancer. Cancer 1997; 80:892.

7. Niwa K, Itoh M, Murase T, Morishita S, Itoh N. Mori H. Tamaya T. Alteration of p53 in ovarian carcinoma clinicopathological correlation and prognostic significance. Br J Cancer 1994; 70:1191.

8. Klemi P-J, Pylkkanen L, Kiilholma P, Kurvinen K, Joensuu H. p53 protein detected by immunohistochemistry as prognostic factor in patients with epithelial ovarian carcinoma. Cancer 1995; 76:1201.

9. Viale G, Maisonneuve P, Bonoldi E, DiBacco A, Bevilacqua P, Panizzoni GA, Radaelli U, Gasparini G. Ann Onc 1997; 8:469.

10. Muto MG, Cramer DW, Tangir J. et al. Frequency of the BRCA1 185delAG muta-

tion among Jewish women with ovarian cancer and matched population controls. Cancer Res 1996; 56:1250.

11. Felip E, Del Campo J, Rubio D, et al. Overexpression of C-erb-B2 in epithelial ovarian cancer. Prognostic value and relationship to response to chemotehrapy. Cancer 1995; 75:2147.

12. Herod JJ, Eliopoulos AG, Warwick J, et al. The prognostic significance of BCL2 and p53 expression in ovarian carcinoma. Cancer Res 1996; 56:2178.

13. Hamada S, Kamada M, Furumoto H, et al. Expression of glutathione S transferase-pi in human ovarian cancer as an indicator of resistance to chemotherapy. Gynecol Oncol 1994; 52:313.

14. Izquierdo MA, van der Zee AGJ, Vermorken JB, et al. Drug resistance associated with Lrp for prediction of response to chemotherapy and prognosis in advanced ovarian cancer. J Natl Cancer Inst 1995; 87:1230.

15. Kavallaris M, Leary JA, Barnett JA, et al. MDR1 and multidrug resistance associated with protein (MRP) gene expression in epithelial ovarian tumours. Cancer Lett 1996; 102:7.

16. Altman DG. Suboptimal analysis using 'optimal' cutpoints. Br J Cancer 1998; 78:556.

17. Burke HB, Henson DE. Specimen banks for prognostic factor research. Arch Path Lab Med 1998; 122:87.

18. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. Cancer 1993; 72:3131.

19. Clark GM, Hilsenbeck SG, Ravdin PM, De Laurentiis M, Osborne CK. Prognostic factors: Rationale and methods of analysis and integration. Breast Cancer Res Treat 1994; 32:105.

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, (eds), Advances in Neural Information Processing Systems 7. Cambridge: MA, MIT Press, 1995:1063.

21. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Pacifistic Grove, CA, Wadsworth and Brooks, 1984.

22. Cox DR. Regression models and life-tables (with discussion). J Royal Stat Soc B 1972:187-220.

23. Breslow NE. Covariance analysis of censored survival data. Biometrics 1974:80.

24. Steel M. Cancer genes: complexes and complexities. Lancet 1993; 342:754.

25. Loomis WF, Sternberg PW. Genetic networks. Science 1995; 269:649.

26. Buratowski S. Mechanisms of gene activation. Science 1995; 270:1773.

27. Sauer F, Hansen SK, Tijan R. Multiple TAFs directing synergistic activation of transcription. Science 1995; 270:1873.

28. Burke HB. Statistical analysis of complex systems in biomedicine. In: Fisher D, Lenz HJ, (eds), Learning From Data: Artificial Intelligence and Statistics V. New York: Springer-Verlag, 1996:251-258.

29. Hebb DO. The Organization of Behavior. New York: Wiley, 1949.

30. Hornik K, Stinchcombe M, White H, Auer P. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Computation 1994; 6:1262.

31. Hornik K, Stinchcombe M, White H. Universal approximation of an unknown function and its derivatives using multilayer feedforward networks. Neural Networks 1990; 3:551.

32. Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995; 346:1135.

33. Dybowski R, Gant V. Artificial neural networks in pathology and medical laboratories. Lancet 1995; 346:1203.

34. Westenskow DR, Orr JA, Simon FH. Intelligent alarms reduce anesthesiologist's response time to critical faults. Anesthesiology 1992; 77:1074.

35. Tourassi GD, Floyd CE, Sostman HD, Coleman RE. Acute pulmonary embolism: artificial neural network approach for diagnosis. Radiology 1993; 189:555.

36. Leong PH, Jabri MA. MATIC - an intracardiac tachycardia classification system. PACE 1992; 15:1317.

37. Gabor AJ, Seyal M. Automated interictal EEG spike detection using artificial neural networks. Electroencephalogr Clin Neurophysiol 1992; 83:271.

38. 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; 86:1853-1859.

39. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr. FE, Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997; 79:857.

40. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995; 346:1075.

41. Swets JA. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. Mahwah, N.J.: Lawrence Erlbaum Associates, 1996.

42. Somer RH. A new asymmetric measure of association for ordinal variables. Am Sociological Rev 1962; 27:799.

43. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psych 1975; 12:387.

44. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982; 143:29.

45. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York, Chapman and Hall, 1993.