1

Integrating Multiple Clinical Tests to Increase Predictive Power

Harry B. Burke

1. Introduction

Clinical tests provide information that can be used by statistical methods to make patient outcome predictions. Outcomes are risk of disease, existence of disease, and prognosis. In this chapter we define and describe predictive factors and clinical prediction and explain how combining predictive factors can increase predictive accuracy, describe the advantages and disadvantages of commonly used statistical methods, and recommend an approach to the reporting of predictive factor research.

2. Predictive 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. For example, the prognostic factor mutant p53 is associated with breast cancer because of its role in the regulation of apoptosis. Such terms as marker, biomarker, predictor, prognosticator, indicator, surrogate factor, and intermediate biomarker have been used to identify variables that are connected to medical outcomes. Their meanings overlap, and their undifferentiated use can cause confusion. All predictive factors are markers of disease (i.e., they are in some way associated with the disease process), but not all markers of disease have sufficient predictive power to be called predictive factors. We 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 model.

Determining whether a marker is a predictive factor requires that:

1. The variable is measured in a defined population;
2. The population is followed until enough outcomes have occurred (i.e., deaths); and
3. The relationship between the variable and the outcome is determined.

From: *Methods in Molecular Medicine, Vol. 14: Tumor Marker Protocols* · c. 1998
Edited by: M. Hanausek and Z. Walaszek © Humana Press Inc., Totowa, NJ

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 predicted outcome always occurs, we say that the predictive factor and the outcome are 100% linked, i.e., the factor has a 100% predictive accuracy *(1)*.

There are three types of predictive factors; risk, diagnostic, and prognostic *(1)*. They differ in their outcomes and predictive power. "Risk" is an ambiguous term. We use "risk" to refer to "risk of disease." "Risk," when used in the context of "risk of recurrence" or "risk of death," is called "probability," as in "probability of recurrence" and "probability of death." Risk factor; the main outcome of interest 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. Diagnostic factor; the main outcome of interest is also incidence of disease. The factor, either alone or in combination with other factors, is close to 100% predictive of disease. Prognostic factor; the main outcome of interest is death. A factor is rarely a strong predictor in isolation from other prognostic factors. There is domain overlap in that risk factors can be prognostic, but they cannot be diagnostic, and diagnostic factors can be prognostic, but they cannot be risk factors.

There are three subtypes of predictive factors: natural history, therapy-dependent, and post-therapy *(1)*. Natural history predictive factors predict the future occurrence (risk), current existence (diagnosis), or course (prognostic) of a disease without an intervention. For risk and prognosis, natural history should the baseline against which all interventions are tested. Therapy-dependent predictive factors assume that there are effective therapies and predict whether the patient will respond to a particular intervention (for example, chemoprevention or chemotherapy). A natural history predictive factor may also be a therapy-dependent predictive factor. Post-therapy predictive factors require that patients respond to an intervention. They predict recurrence of the risk of disease or recurrence of the disease.

The predictive power of a factor depends on its intrinsic and extrinsic powers. The intrinsic predictive power of a factor is related to its "connectedness" to the disease process, i.e., its association to the disease process. 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 it (i.e., a secondary infection). 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 its relationship to the other factors (degree of predictive overlap);
2. How easy it is to collect and measure the factor; and
3. The degree to which the selected statistical method is able to capture the individual factor's predictive information and to integrate it with the information of other 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. The usual strategy, when dealing with predictive factors, is to combine several in a predictive model. The most useful grouping of factors is one in which all of the factors are powerful and predictively orthogonal to each other, i.e., they index independent aspects of the disease process. If they represent aspects of the disease that are not independent of each other, then to the degree that their information overlaps is the degree to which one will not add predictive power. The statistical method employed must be able to capture the complexity of the disease process indexed by the predictive factors.

A predictive model for a specific outcome is the result of entering one or more predictive factors into a statistical method. The statistical method attempts to capture the relationship between the factors and the outcome. For example, the mathematical formula generated by the logistic regression statistical method relates the predictive factors (input variables), in terms of their β-coefficients, to a binary disease outcome (relapse, death, and so forth). It should be noted that the predictive power of a factor depends on the specific statistical method selected and on the other factors selected to be included in the model. The statistical model that results from the application of a statistical method, learning the relationship between the factors and the outcome, may or may not be the most efficient at capturing the predictive power of the factors.

Before discussing specific statistical methods, it is important to distinguish among significance, accuracy, and importance *(2)*. Model significance asks if the observed predictions are really different from those produced by another model or from those resulting from chance.

Significance is not accuracy. Accuracy is the association between the model's predictions and the known outcomes in a test population. The importance of a model or a factor is determined by whether the model or factor possesses sufficient accuracy to be useful in answering a particular clinical question. Finally, the assessment of model or factor significance, accuracy, and importance must be based on test data set results, not on training data set results.

3. Advantages and Disadvantages of Statistical Methods

Many methods can be used to combine predictive factors. In cancer, they include bins, stages, and indexes; decision trees; and regression methods, including logistic, proportional hazards, and artificial neural networks.

Bins are the result of the mutually exclusive and exhaustive partitioning of discrete variables. Each combination of variable values is a bin, and all patients are placed in the bin corresponding to their variable value combination *(2)*. An example is the TNM classification of breast cancer *(3)*. Tumor size (Tis, T1, T2, T3, T4), number of positive regional lymph nodes (N0, N1, N2, N3), and existence of metastases (M0, M1) produce 40 bins *(2)*.

Each patient in a bin receives the same prediction; namely, the most frequent outcome. If there are enough patients in each bin, it can be shown that the most frequent outcome is the best predictor of the true outcome. In other words, no prediction model can be more accurate than a bin model if the variables are discrete and the population is large. Problems with bin models *(2)* include:

1. Continuous variables must be cut up into discrete variables. This almost always results in a loss of predictive information and therefore a loss of accuracy.
2. As the number of discrete variables increases, the number of bins increases exponentially. In order to maintain accuracy, there must be a corresponding exponential increase in the size of the patient population.
3. The proliferation of bins reduces the ability to understand the phenomena. Bin proliferation negates the main advantage of a bin model; namely, its ease of understanding and ease of use.

Bin models are rarely used in situations in which there are more than two or three predictive factors or where each factor possesses more than a few strata.

A partial solution to the problems of a bin model is a stage model *(2)*. A stage model is the grouping of bins into super-bins. The justification for the grouping is the assumption that the factors selected represent "stages" of the disease process. For example, in breast cancer, the TNM staging system combines 40 TNM classification bins into six super-bins (TNM stages) based on decreasing survival ("stages of survival").

A small set of stages has the potential to maintain explanatory simplicity and ease of use. Problems with stage models include:

1. The combining of bins into super-bins/stages can substantially reduce predictive accuracy.
2. Stage systems do not overcome the exponential increase in bins and patients associated with adding a variable to the analysis: They just delay the problem at a cost in predictive accuracy. If the stages are held constant when variables (and their associated bins) are added to 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 cutting up continuous variables, with the resulting loss in predictive accuracy, remains.
4. Finally, if a single staging system is used for more than one cancer site, the staging rules may be more applicable to some sites than to other sites. The sites to which they do not apply will experience major losses in predictive accuracy.

Indexes associate numerical scores (usually based on a bounded, linear scale) with bins or groups of bins. Each score is associated with one of a small number of disease stages (usually a severity of illness system). Each patient receives the prediction of the stage in which their score places them. Indexes offer some flexibility in the grouping of bins, but at the cost of further degradation in predictive accuracy because additional information is lost. The simplest example of an index is the Apgar. An example in breast cancer is the Nottingham Index *(4)*.

The accuracy of different stratifications of a predictive factor(s) can be compared. For a specific site (i.e., breast) and predictor(s) (tumor size <2, 2–5, >5) any bin or group of bins, or stage (bin or index) or group of stages, can be compared, in terms of a specific outcome, with another stratification (tumor size <1, 1–<2, 2–<3, 3–<4, 4–<5, 5–>5). This contrast can be over a single time interval without respect to events within the interval (i.e., logistic regression) or with respect to the events within the interval *(5,6)*. For a single interval without respect to events within the interval, accuracy has been assessed by several discriminative association approaches, including Goodman and Kruskall's Gamma *(7)*, Kendall's Tau *(8)*, or the area under the receiver operating characteristic *(9)*.

The usual descriptive approach for contrasting predictive factors across a series of event time intervals is the Kaplan-Meier product-limit method *(5)* (inferential methods that can accommodate continuous variables, and that usually assume proportional hazards, will be discussed later when regression methods are presented). A Kaplan-Meier plot should always include confidence intervals for each stratum (i.e., each step function). A significant difference within a Kaplan-Meier stratification (tumor size <2, 2–5, >5) is usually assessed by a log-rank test *(10)*. It is important to note that there is currently no method for comparing the accuracy of two different Kaplan-Meier plots (i.e., two different stratifications of the same predictive factors). It is incorrect to use the *p*-value of the log-rank test to select one stratification over another, because the log-rank test only determines whether a stratification is likely to have occurred by chance. An extreme stratification may result in smaller *p*-values, but it may also reduce predictive accuracy.

Decision trees split predictive factors to maximize predictive power using a loss function, such as the log-likelihood and a greedy search algorithm. A well-known decision tree approach is the Classification and Regression Trees (CART) recursive partitioning method *(11)*. Empirically, we have not found CART, either pruned or shrunk, 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 when there are more than two or three splits.

Univariate regression methods are not appropriate for determining whether a variable is a predictive factor. Univariate methods should not be used, because new variables must be assessed in the context of the known factors, and because some variables are only predictive when they interact with another variable.

Logistic regression assess 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 binary outcome problems (i.e., recurrence and survival). Its limitation is that it usually spans a large time interval and does not distinguish when events occur within the time interval. This limitation can be overcome if several sub-time intervals are created within the overall time interval. Logistic regression models can be created for each sub-time interval. Censoring can be accommodated by removing cases that are censored within the time interval that censoring occurs.

Proportional hazards methods include the Cox *(6)* and less commonly the Weibull or exponential *(12)*. Proportional hazards methods assume that the hazard of each patient is proportional to the hazards of all the other patients, and that a patient's hazard is related to that patient's relative risk. The Cox model does not create survival curves. For Cox-related survival curves, a baseline hazard must be introduced (for example, Breslow-Cox estimates) *(13)*. Some researchers incorrectly believe that the Cox is the only regression method that can deal with censoring (*see* paragraph on logistic regression above). Because, in cancer, the proportional hazards' assumption may be violated, researchers who use the Cox model must demonstrate that the proportional hazards assumption holds for their population.

Artificial neural networks are a general regression method *(14,15)*. They can perform almost any regression task. In addition, three-layer artificial neural networks automatically capture nonlinearity and complex interactions. They can handle censoring in the same way that multi-interval logistic regression handles censoring. Artificial neural networks are as transparent as the phenomena contained in the data. For simple phenomena, artificial neural networks are easily understood; for complex phenomena they are complex and less easily understood. Artificial neural networks are especially recommended in the domain of complex systems (e.g., the molecular-genetic domain of cancer).

4. Reporting Predictive Factor Research Results

There is a great deal of variation in the reporting of predictive factor results. This variability makes it difficult to understand and compare results. The following is a recommended approach to reporting the discovery of a new predictive factor or the validation of an existing factor.

For a defined subset of patients with the ___a___ disease, ___b___ is a ___c___ predictive factor for ___d___ when assayed ___e___ by ___f___, for the ___g___ on a test data set with ___h___ characteristics, the ___i___ is significant at the ___j___ level using the ___k___ statistical method, which also incorporates ___l___ predictive factors, for ___m___ therapy. Using the ___n___ method to assess its accuracy, the ___k___ statistical model is ___o___ accurate on the test data set.

"Defined" means specification of collection method, inclusion and exclusion criteria, and so forth.

a: Name of disease.
b: Name of the predictive factor.
c: Type and subtype of predictive factor (i.e., risk, diagnosis, prognosis; natural history, therapy-dependent, post-therapy).
d: Outcome (i.e., 5-yr breast cancer-specific survival).
e: Time of assay (i.e., at discovery, prior to therapy, after therapy).
f: Specific laboratory method (i.e., immunohistochemistry).
g: If stratified, the specific range/cut-point/and so forth of the prognostic factor. If the variable value is based on rater judgment, then Cohen's κ should be reported.
h: Relevant characteristics of the data set, including:
  1. Data set size,
  2. Number of events, and
  3. Whether the therapy was randomized.
i: The value and confidence interval.
j: For example, $p < 0.05$ for one test of the data. If multiple tests of the data are performed, an adjustment may be required.
k: Type of multivariate statistical method (i.e., logistic regression, Cox).
l: Other relevant prognostic factors, if they are included in the multivariate model.
m: Specific type of surgery, chemotherapy, radiation therapy.
n: Area under the receiver operating characteristic (Az) R^2, χ-square, etc.
o: Numerical value and its range of possible values (i.e., Az = 0.75, 0.50, –1.0).

References

1. Burke, H. B. (1994) Increasing the power of surrogate endpoint biomarkers: aggregation of predictive factors. *J. Cell. Biochem.* **19,** 278–282.
2. Burke, H. B. and Henson, D. H. (1993) Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* **72,** 3131–3135.

3. Beahrs, O. H., Henson, D. E., Hutter, R. V. P., and Kennedy, B. J. (1992) *Manual for Staging of Cancer,* 4th ed., Lippincott, Philadelphia, PA.
4. Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., Nicholson, R. I., and Griffiths, K. (1982) A prognostic index in primary breast cancer. *Br. J. Cancer* **45,** 361–366.
5. Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53,** 457–481.
6. Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. Royal Stat. Soc. B.,* pp. 187–220.
7. Goodman, L. A. and Kruskal, W. H. (1954) Measures of association for cross classifications. *J. Am. Stat. Assoc.* **49,** 732–764.
8. Kendall, M. G. (1962) *Rank Correlation Methods.* Hafner, New York.
9. Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psy.* **12,** 387–415.
10. Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50,** 163–170.
11. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth and Brooks, Pacific Grove, CA.
12. Evans, M., Hastings, N., and Peacock, B. (1993) *Statistical Distributions,* 2nd ed., Wiley, New York.
13. Breslow, N. E. (1974) Covariance analysis of censored survival data. *Biometrics* **30,** 80–99.
14. Burke, H. B. (1994) Artificial neural networks for cancer research: outcome prediction. *Sem. Surg. Onc.* **10,** 73–79.
15. Burke, H. B., Rosen, D. B., and Goodman, P. H. (1995) Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival, in *Advances in Neural Information Processing Systems*, vol. 7 (Tesauro, G., Touretzky, D. S., Leen, T. K., eds.), MIT Press, Cambridge, MA, pp. 1063–1067.