EDITORIALS

Breast Cancer Prognostic Factors: Evaluation Guidelines

*William L. McGuire**

In the present issue of this journal, Thor and colleagues attempt to evaluate heat shock protein 27 (also known as the 27 000-dalton stress response protein or srp-27) measured in tumors from breast cancer patients as a prognostic factor (*1*). They report significant correlations between srp-27 overexpression and other measured prognostic factors as well as between srp-27 overexpression and a shorter disease-free survival period. However, they state that a multivariate analysis failed to recognize srp-27 expression as a significant independent predictive factor. Before commenting on the Thor paper, it might be useful to consider first the broader problem of how the casual reader should evaluate papers dealing with prognostic factors in breast cancer. The use of prognostic factors to help select breast cancer patients for adjuvant therapy is of considerable concern to the oncology community (*2*). This need for selection of prognostically less favorable cases is stimulating investigators to identify new and more powerful prognostic factors. Unfortunately, however, this identification process is becoming more confusing because of a lack of guidelines for investigators to use to study new factors and for reviewers and readers to use to evaluate papers on this topic. Listed here are the minimal criteria that must be considered when one is attempting to evaluate a new prognostic factor.

Prognostic Factor Evaluation Guidelines

1. Biological hypothesis
2. Pilot study vs. definitive study
3. Sample size calculation
4. Patient population bias
5. Methodological validation
6. Optimized cutoff values
7. Reproducibility

First, the factor to be studied should possess clear biological significance, and the investigator(s) should clearly define the hypothesis to be tested. Next, the investigator(s) should state whether their study is a pilot study, a definitive study, or a confirmatory study. If the author states that a given study is only a pilot study, then evaluation rules different from those used when a study is a definitive study may apply. For example, the number of patients in a pilot study may be quite small, or the statistical significance may be found in only a restricted subset of the patients studied. If a study is labeled as a pilot study, one should not attempt to draw conclusive clinical implications from its results. A pilot study should only provide a clue for the next step, which is a definitive study. The definitive study should be the solid basis for evaluating a prognostic factor. Unfortunately, it is too often poorly designed and considers too few patients. Consequently, it becomes only another pilot study of the same prognostic factor. Examples of such studies are commonplace in the breast cancer prognostic factor field. The third type of study is one designed to confirm a definitive study. Such confirmation is extremely important and necessary to ascertain that a particular assay works well elsewhere in a different set of patients.

The next consideration is the requirement of an adequate sample size for meaningful calculations in a definitive study. It is paradoxical that we readily accept studies with only 50 to 150 patients to define the importance of prognostic factors in predicting disease-free survival and accept the conclusions from these studies as definitive. Yet, if we were studying the efficacy of a drug treatment on influencing disease-free survival, we would insist on a formal sample size calculation and reject breast cancer studies with only 50 to 150 patients. Given certain information, statisticians should quickly be able to determine the number of patients required to evaluate a given prognostic factor. The information needed will include an estimate of the magnitude of the contribution of that factor, the extent of correlation of that factor with other prognostic factors, the distribution of that factor in the population, the length of followup, and the number of recurrences and deaths in the sample population. One purpose of the pilot study is to generate the estimates above so that a final definitive study can be designed.

In a definitive prognostic factor study, the patient population must be appropriate for the hypothesis and carefully scrutinized for intrinsic biases. If the hypothesis or pilot study suggests an effect in axillary node-negative patients, the definitive study should be designed entirely with axillary node-negative patients. Often, node-positive patients or more advanced-stage patients are included to give the appearance of large numbers, whereas in fact they may not be contributing to the test of the hypothesis. Also, the effect of treatment on the patient population needs to be considered and shown not to be a confounding factor in evaluating a prognostic factor.

Patient population bias may result from retrospective tumor bank studies. For example, since disproportionate numbers of larger tumors end up in frozen tumor banks, the results from such a study could generalize only to larger tumors. To make the results more representative of the breast cancer population as a whole, one might use stratified sampling with sampling fractions based on the tumor size distribution of large tumor registry data bases. Population bias may also result from studying only those patients entered into clinical trials. Again, the smallest tumors with the most favorable prognosis might be underrepresented. Although

Received December 20, 1990; accepted December 21, 1990.

**Correspondence to:* William L. McGuire, MD, Department of Medical Oncology, University of Texas Health Science Center, 7703 Floyd Curl Dr, San Antonio, TX 78284-7884.

unproven, there is a feeling prevailing among oncologists that patients with favorable clinical characteristics are less frequently entered into clinical trials than patients with a poorer prognosis. This practice would result in disease-free survival figures associated with the presence or absence of a prognostic factor that are worse than in the general population.

Methodological validation, too, is frequently overlooked. For example, if blot techniques are used to measure DNA, RNA, or protein factors, how does one know whether a negative value is due to the absence or paucity of tumor cells in the specimen homogenized for the assay? We should insist that adjacent sections be reviewed histologically to assure that a minimum number of tumor cells are assayed. Immunohistochemical assays, on the other hand, address the problem of heterogeneity but are fraught with sensitivity and specificity problems. Increasing the dilution (ie, decreasing the concentration) of antibody decreases sensitivity, while decreasing the dilution (ie, increasing the concentration) results in nonspecific immunostaining. In contrast to electrophoretic blot techniques in which a molecule of a particular size is being measured, several different molecules could unknowingly be measured by immunostaining (due to cross-reactivity or nonspecific staining). Therefore, we should insist upon parallel demonstration by Western blot or immunoprecipitation techniques of the protein being measured in immunostaining studies. Since different assay methods have different sensitivities, the sensitivity used in a particular study must be defined. For a confirmation study it is essential that the sensitivity of the assay used be comparable to that used in the earlier study.

Apart from the sensitivity of the assay, which defines a positive or negative chemical value, we usually assign a "clinical cutoff value" to separate high from low or overexpression from normal expression, etc. These cutoff points may be arbitrarily assigned as the median value or a particular percentile. From a biological point of view, such an assignment might be quite inappropriate. Alternatively, we may sequentially examine every conceivable cutoff value to maximize the separation of disease-free survival curves. This procedure makes good sense, particularly when the final assay result must be dichotomized as yes/no, high/low, etc. The criticism of this approach is that if the total data set is used to find the optimal cutoff value, one is then unable to validate the particular choice of the cutoff value in an independent data set. If optimized cutoff values are to be used, we should insist that a "training" data set be used to determine the cutoff value and that an independent "test" data set be used to validate the choice.

Finally, the result of a particular study or assay in a particular medical research center must be readily reproduced elsewhere if it is to have clinical usefulness. The design of a confirmatory study should be the same as the definitive study that it attempts to duplicate. This replication of study design is rare in the prognostic factor field. Often completely different assays or experimental designs are used without attempts to standardize the laboratory results or clinical outcomes. Even worse, investigators may study fewer patients than studied originally in the definitive study and then conclude that they cannot confirm the results of the original definitive study! The prognostic factor literature is already satiated with studies that were intended to be either definitive or confirmatory of definitive studies but instead resulted in a failure to meet the requirements of either a definitive or a confirmatory study for the reasons stated above. The time has come to establish guidelines for investigators to follow and reviewers and readers to use in evaluating such efforts.

How do these evaluation guidelines apply to the paper by Thor and colleagues on srp-27? Although it is tempting to use this paper to illustrate the problems raised by the evaluation criteria above, such an endeavor probably would not be fair at this time, since these guidelines are newly proposed; they require refinement and would benefit from the consideration and input of others. Furthermore, there are few prognostic papers currently in the literature that would satisfy the proposed guidelines. Nevertheless, I do agree with the concluding comments of the authors that "additional studies are needed to discern both the biological contributions of srp-27 and how this marker may potentially influence the clinical behavior of breast cancers." I would add strongly that these additional studies should follow the guidelines outlined herein.

References

(1) Thor A, Benz C, Moore D II, et al: Stress-response protein (srp-27) determination in primary human breast carcinomas: Clinical, histologic, and prognostic correlations. J Natl Cancer Inst 83:170–178, 1991

(2) McGuire WL, Tandon AK, Allred DC, et al: How to use prognostic factors in axillary node-negative breast cancer patients. J Natl Cancer Inst 82:1006–1015, 1990