Arbiter:

D.F. Hayes*

Lombardi Cancer Center, Georgetown University Medical Center, NRB E504, 3970 Reservoir Rd NW, Washington, D.C. 20007, USA

Do we need more prognostic factors in node-negative breast cancer? As I read the accompanying opinion pieces by two highly regarded experts in the field regarding this topic, I was struck by a curious observation. Although they were supposed to take 'pro' or 'con' positions, in reality both authors came to the same conclusion. Their comments can be paraphrased in the following statement: 'Yes, we do need new prognostic factors, but they need to be more precise and reliable'.

Both Drs Thomssen and Janicke and Drs Kaufmann and Scharl refer indirectly to the chaos that has existed in the field of tumour markers, which has led to near paralysis in trying to select markers that might have clinical utility. Indeed, 5 years ago, a panel of experts was convened by the American Society of Clinical Oncology (ASCO) to establish guidelines for the use of both tissue-based and circulating tumour markers in patients with breast cancer. The published guidelines from this Expert Panel regarding tissue-based markers were quite conservative, recommending only the use of oestrogen receptor (ER) and progesterone receptor (PR) to determine the likelihood of benefit from endocrine therapy [1,2]. In spite of nearly 20 years of remarkable advances stemming from the revolution in molecular biology, the members of the Expert Panel were unable to determine whether any other markers could actually be used to reliably guide patient care in a reproducible fashion.

From the comments by Drs Thomssen, Janicke, Kaufmannn and Scharl, and the recommendations of the last St Gallen Conference, it appears that a similarly conservative approach is taken in Europe [3]. Why haven't we incorporated new markers, such as p53 abnormalities, erbB-2 amplification/overexpression, angiogenesis, and markers of tissue infiltration such as uPA and PAI-1 into routine clinical care? I believe there are two reasons for our lack of acceptance of new markers: (1) There has been no consensus structure in which to perform tumour marker studies and, therefore, there are no consistent methods to design or evaluate tumour marker investigations; and (2) The decisions to treat or not to treat a patient with adjuvant systemic therapy are based on complex perceptions regarding the relative benefits and the relative costs and toxicities. These differ depending on the agent, the patient, the caregiver and the society in which they exist.

1. Tumour Marker Utility Grading System (TMUGS)

I will consider these issues separately. When the ASCO Tumour Marker Expert Panel was convened in 1995, the members were struck by the remarkable heterogeneity among tumour marker studies, even those ostensibly addressing the same issue. An analogy (albeit a negative one) can be drawn from the state-of-the-art strategies for performing therapeutic trials 30–40 years ago. At that time, to impose some order on the results that would be derived from clinical trials, groups of investigators devised clinical trial designs (phase I, phase II and phase III) and scales for analysis (toxicity scales, performance status scales and response criteria) that, with some modifications, persist and serve us well today [4]. Importantly, such trials demand a prospective hypothesis and a preset 'protocol' that dictates the methods to be followed in conducting the study and analysing the data. In fact, these clinical investigators were applying the scientific method, which had been well established in laboratory science, to the clinical setting. Even using such strict methods, we still often require meta-analyses of the results from several well designed but relatively underpowered clinical trials to determine how beneficial a specific therapy really is [5,6]. Indeed, any investigator who claims that his or her single, retrospectively performed tumour marker study demonstrates definitive clinical utility should review Fig. 1 from the most recently published report of the Oxford overview of randomised trials of polychemotherapy for breast cancer [6]. Although the overall data confirm the benefits of adjuvant chemotherapy, there are several prospective randomised clinical trials that show either no benefit for treatment, or even that treatment results in worse outcomes.

With certain exceptions, no such system has been devised for tumour marker studies [7,8]. Rather, most tumour marker investigations have been performed because an investigator has access to an interesting assay and happens to have a group of available patient specimens. Rarely is there quality control over how the

* Tel.: +1-202-687-3013; fax: +1-202-687-0338.
E-mail address: hayesdf@gunet.georgetown.edu (D.F. Hayes).

PII: S0959-8049(99)00303-2

patients were selected, how the specimens were collected, processed or stored, or how the assay was performed. Moreover, tumour marker studies almost never include prospective power calculations or analytical plans. Whilst results from such studies are adequate to generate hypotheses, we should not be surprised that the results that they generate are poorly reproducible and generally unreliable.

As a function of trying to make order from the chaos of available tumour marker studies, several members of the ASCO Tumour Marker Expert Panel developed, and published, a Tumour Marker Utility Grading System (TMUGS) [9]. TMUGS helps a reviewer organise available studies with regard to the assay technology used to evaluate a specific marker. It also provides a list of possible uses for which the marker might be applied, recommends clinical outcomes that should be influenced by the results of a marker, and suggests a semi-quantitative scale of utility grades (0–3+) that the reviewer might assign to the marker. Only those markers that are assigned utility grades of 2+ or 3+ would be considered satisfactory for recommendation in routine clinical practice. Perhaps most importantly, a critical feature of TMUGS is a definition of the quality of the evidence available to support assignation of a utility grade. Ideally, one would have Level of Evidence (LOE) I data to assign a utility grade to a given marker for a particular use. LOE I studies consist of either a highly powered prospective randomised trial in which the marker is the primary objective of the study or a critically performed meta-analysis of lesser LOE studies. LOE II studies include those in which the marker question is a secondary objective within a prospective clinical trial that is performed to address a therapeutic question. LOE III studies consist of the hypothesis-generating investigations that are represented by most currently available tumour marker publications.

2. TMUGS 'plus'

Even if LOE I studies are available, how does a clinician decide when to use the marker in routine clinical practice? Too often, investigators observe that a group of patients that are 'positive' for a given marker have an outcome that is statistically significantly different from another group that is 'negative'. Therefore, they conclude that the marker should become part of routine evaluation of patients with the disease (as illustrated by a recent paper by the International Breast Cancer Study Group [10]). Two important components of decision making are often lost in these conclusions: (1) the difference between prognosis and prediction; and (2) the difference between being statistically significant and clinically useful. Drs Thomssen, Janicke, Kaufmann and Scharl refer to the simple but groundbreaking definitions of prognosis and prediction, as first proposed by McGuire and later by Gasparini and colleagues [8,11]. However, very little attention has been paid to the magnitude of difference in outcomes between the two groups (those that are positive for a factor versus those that are negative), whether independent ('prognostic') or dependent ('predictive') of specific therapies. The utility of a marker depends on both this magnitude of difference and whether the difference is unlikely to be due to chance alone.

These concepts are illustrated in Figs. 1 and 2. Fig. 1(a) represents an example of two pure prognostic factors, whilst Fig. 1(b) represents two pure predictive factors. The example provided in Fig. 1(a) represents poor prognostic factors; that is, 'factor-positive' patients have a poorer clinical outcome than 'factor-negative' patients. However, the poor prognosis is independent of therapy. Therefore, both groups may benefit from therapy equally, but factor-positive patients still have a worse outcome than factor-negative patients do. Note that the distance between the factor-'negative' line and the factor-'positive' lines represents the relative magnitude of the prognostic strength of each factor. In this example, as is often the case in published markers of tumour markers, the difference between the factor-negative and -positive lines for both markers has been statistically established, with P values less than 0.05. In

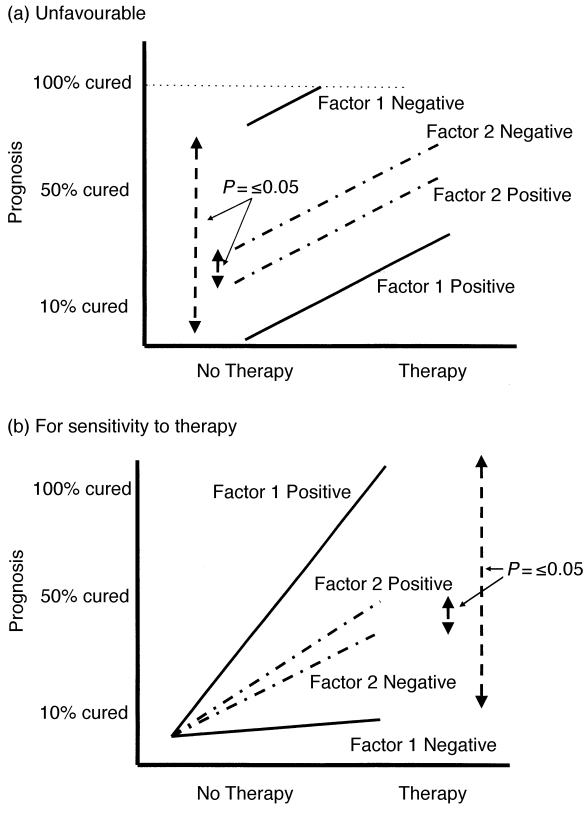

Fig. 1. Schematic examples of pure prognostic (a) and pure predictive (b) factors. Strong factor (1). Weak factor (2).

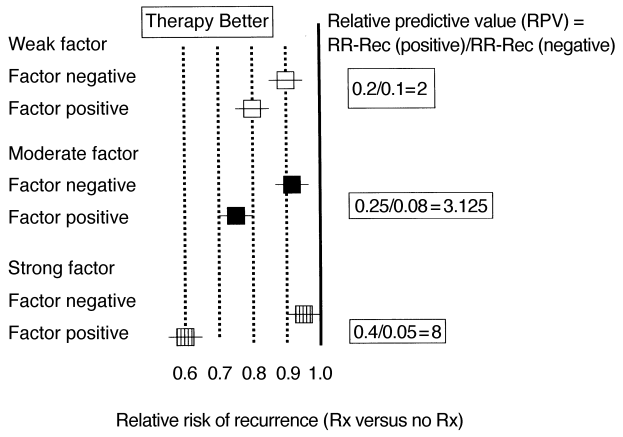

Fig. 2. Schematic example of method to determine relative predictive value (RPV) from results of prospective randomised clinical trial in which patients are randomly assigned to therapy versus no therapy.

other words, the difference in outcomes between factor-positive and -negative patients is unlikely to be due to chance alone for either the strong or the weak factor.

However, the clinical utility of these two factors (strong and weak) is quite different. There is a huge separation between the factor-negative and positive lines for the strong factor. Indeed, the outcome for the factor-negative patients is so good that, depending on the toxicities and costs of therapy, we might decide not to recommend adjuvant systemic therapy. In contrast, although statistically different, the outcomes for two groups of patients divided by a weak factor are not sufficiently substantial to be clinically important. In other words, we would probably offer the same treatment recommendations to both groups of patients. Therefore, although important to ensure that the investigator's observation is not due to chance alone, the *P* value does not determine clinical utility. Rather, it is the magnitude of the relative risk of poor outcome between factor-negative and -positive patients that determines whether the factor is useful to the clinician and patient.

Similar considerations can be applied to a predictive factor, as illustrated in Fig. 1(b). In this example, the illustrated markers are positive predictive factors for response to a specific therapy, and, again, illustrations are provided for strong and weak factors. Note that in the absence of the specific therapy ('No Therapy' in Fig. 1), a pure predictive factor does not separate the population into two patient groups. It is only in the presence of the specific therapy for which the factor is predictive that the outcomes of 'positive' and 'negative' patients differ.

Two final considerations must be addressed. Very few factors are absolutely accurate. If one is using such factors to determine whether to not treat a patient, then one has to be willing to accept that the patient may forego some potential benefit. If a patient is willing to accept all toxicities and costs to achieve any possible benefit (even as low as 1%), then tumour markers should not be used. Placed in the context of node-negative patients for whom one is considering therapy, no factors (prognostic or predictive) are likely to identify a group of patients for whom adjuvant systemic therapy does not offer some chance of benefit. Thus, one also has to consider the cost of the specific therapy, both with regard to toxicities and financial expense. In this regard, a patient might elect to take tamoxifen as adjuvant systemic therapy even if her prognosis is very good (for example, a 1 cm, well differentiated cancer) and her odds of benefiting are very low (for example, if her tumour is ER poor). Even in these circumstances, an occasional patient will avoid an incurable recurrence by taking tamoxifen. Is this small potential benefit worth the exposure to the relatively low incidence of potentially lethal toxicities, the more common but usually tolerable side-effects, and the modest cost of 5 years of tamoxifen? Some individuals might say yes, others no [12]. Moreover, some healthcare systems might be willing and able to shoulder the burden of such an expense, whilst others might demand a much higher benefit/risk ratio to justify the financial cost. It is likely that these deliberations will not only differ from patient-to-patient and doctor-to-doctor, but between different national healthcare delivery systems, as well.

The deliberations regarding chemotherapy are influenced by the higher rate of therapy-related mortality (as high as 1–2%) in some adjuvant studies and the dramatically increased risk of intolerable side-effects, as well as the considerable economic cost over a short period of time (months, rather than over 5 years with tamoxifen). In this case, patients, physicians and third-party payers might choose therapy only if their odds of benefiting are substantially higher. Moreover, several active chemotherapeutic agents are now available, with proven efficacy in the adjuvant setting but with considerably different toxicity profiles. Predictive factors that permit selection of specific agents for individual patients could be very important.

Thus, a new prognostic factor is unlikely to be of much benefit regarding endocrine therapy, unless it separates patients into those absolutely unlikely to recur (and, therefore, cannot benefit), or those ER-positive patients that are absolutely unlikely to respond. Otherwise, most patients and caregivers seem likely to elect treatment. With the increasing experience of delivering chemotherapy in shorter courses, the availability of newer anti-emetics, and the current lack of emergence of important long-term complications related to adjuvant chemotherapy, one wonders whether the same is true for chemotherapy? None the less, even 3 months of chemotherapy has a substantial impact on a patient's quality of life, and factors that reliably place a patient into an extraordinarily good prognostic category would be of great value for most.

Therefore, it is reasonable to place prognostic and predictive factors into categories, for example 'weak', 'moderate' and 'strong' [13]. The prognostic strengths of different factors can be determined by comparing outcomes of factor-positive versus factor-negative patients in a non-randomised trial, but the effects of therapy must be considered. In this regard, studies that use randomly collected patient samples represent LOE III investigations and are unlikely to provide definitive information for routine clinical application.

In contrast, the predictive strength of a factor can only be determined by comparing relative outcomes in factor-negative and -positive patients treated with the therapy of interest versus the relative outcomes of factor-negative and -positive patients who did not receive the therapy [13]. Preferably, such a comparison should be performed in the context of a prospective, randomised clinical trial. Fig. 2 is a conceptual illustration of how one might estimate the predictive strength from the results of a randomised trial or an overview of several such trials. In this example, the relative predictive strengths of weak, moderate and strong factors are illustrated. One can estimate the relative predictive value (RPV) of a marker by dividing the relative risk (RR) reduction for recurrence for treated patients versus untreated patients who are factor-positive by the RR reduction for recurrence for factor-negative patients. For example, in Fig. 2, consider the effects of therapy versus no therapy for a population of patients divided by a weak predictive factor. The RR for recurrence for factor-negative patients is 0.9 (in other words, they only experienced a 10% proportional reduction in recurrence because of treatment). Weak factor-positive patients who are treated have a relative risk for recurrence of 0.8, compared with those who are not treated. Therefore, the relative predictive value (RPV) for this factor is $0.2/0.1 = 2$. Using the example of a moderately strong predictive factor, positive patients benefit more from treatment, with a RR of recurrence compared with untreated patients of 0.75, whilst negative patients benefit less from treatment ($RR = 0.92$). The calculated RPV for this factor in regards to this therapy is $0.25/0.08 = 3.125$. Finally, in this example, a strong predictive factor divides the population into those with a 40% or greater proportional reduction if they are factor positive (RR for recurrence for treated versus untreated = 0.6) and those factor-negative patients with a very low chance of benefit (RR for recurrence = 0.95). In this case, the $RPV = 0.4/0.05 = 8$. We have previously suggested that predictive factors with RPV of 1–2 be considered weak, those with RPV of 2–4 be considered moderate, and those with RPV of >4 be considered strong [13]. For routine clinical purposes, we would reject the former, accept the latter, and consider the use of the middle group with hesitation.

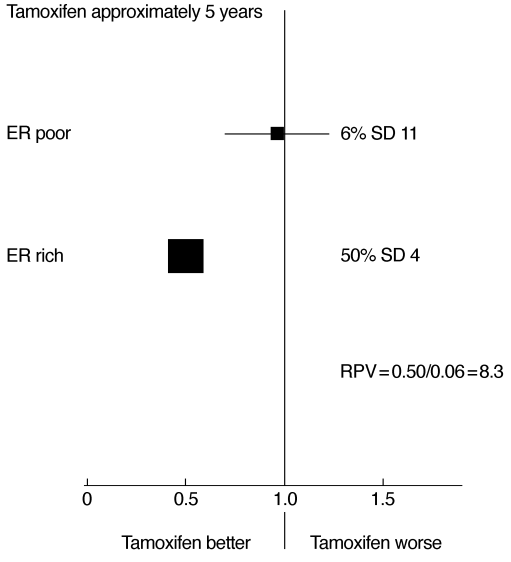

Fig. 3. Relative predictive value of oestrogen receptor for use of tamoxifen as adjuvant therapy for 5 years. ER, oestrogen receptor; SD, standard deviation; RPV, relative predicted value. Modified from [5] with permission.

Tumour ER content as a predictive factor for tamoxifen in the adjuvant setting is the best real-life example in breast cancer, and perhaps in all of oncology. Although it took nearly 20 years to compile, the Oxford Overview now permits us to reasonably estimate the relative strength of ER as a predictive marker for the benefits from tamoxifen [5]. From Fig. 3, one can see that the reduction of recurrence or mortality due to adjuvant tamoxifen in ER-rich versus ER-poor patients differs substantially. The risk reduction for recurrence in tamoxifen-treated ER-poor patients compared with untreated patients is 6% (SD11) ($RR = 0.94$). In contrast, the relative risk reduction in ER-rich patients who received 5 years of tamoxifen is 50% (SD4) or more ($RR = 0.50$). Therefore, we can estimate that the RPV of ER for tamoxifen in this setting is $0.50/0.06 = 8.3$, and we can safely conclude that ER is a very strong predictive factor.

In summary, YES — we do need new prognostic and predictive factors for patients with newly diagnosed breast cancer! However, as implied by Drs Thomssen, Janicke, Kaufmann and Scharl, we do not need weak ones, and we need to be able to estimate reliably the strength of new factors from adequately generated scientific evidence. LOE III studies are still important, but they should be used to generate new hypotheses, not to determine clinical utility. Once generated, these hypotheses should be prospectively tested in well designed clinical investigations. I congratulate Dr Janicke and his colleagues in Germany for their visionary prospective trial testing whether uPA and PAI-1 are adequate to use for clinical decision making. The debate regarding uPA and PAI-1 has been ongoing for nearly

10 years. This LOE I trial should at least give the clinician sufficient information to place the relative prognostic strength of the markers in the context of toxicities of therapy, individual patient wishes, and societal priorities. We need this type of study so that we can make intelligent clinical decisions to help our patients as they face this terrible disease.

References

1. ASCO Expert Panel. Clinical practice guidelines for the use of tumor markers in breast and colorectal cancer: report of the American Society of Clinical Oncology Expert Panel. *J Clin Oncol* 1996, **14**, 2843–2877.
2. ASCO Expert Panel. 1997 update of recommendations for the use of tumor markers in breast and colorectal cancer. *J Clin Oncol* 1997, **16**, 793–795.
3. Goldhirsch A, Glick JH, Gelber RD, Senn HJ. Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. *J Natl Cancer Inst* 1998, **90**, 1601–1608.
4. Simon R. Design and conduct of clinical trials. In DeVita VT Jr, Hellman S, Rosenberg SA, eds. *Cancer: principles and practice of oncology, 4th edn*. Philadelphia, J.B. Lippincott Co, 1993, 418–440.
5. Early Breast Cancer Trialist's Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* 1998, **351**, 1451–1467.
6. Early Breast Cancer Trialist's Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomized trials. *Lancet* 1998, **352**, 930–942.
7. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994, **69**, 979–985.
8. McGuire WL. Breast cancer prognostic factors: evaluation guidelines. *J Natl Cancer Inst* 1991, **83**, 154–155.
9. Hayes DF, Bast R, Desch CE, *et al*. A tumor marker utility grading system (TMUGS): a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst* 1996, **88**, 1456–1466.
10. Cote RJ, Peterson HF, Chaiwun B, *et al*. Role of immunohistochemical detection of lymph-node metastases in management of breast cancer. International Breast Cancer Study Group. *Lancet* 1999, **354**, 896–900.
11. Gasparini G, Pozza F, Harris AL. Evaluating the potential usefulness of new prognostic and predictive indicators in node-negative breast cancer patients. *J Natl Cancer Inst*, 1993, **85**, 1206–1219.
12. Ravdin P, Siminoff I, Harvey J. Survey of breast cancer patients concerning their knowledge and expectations of adjuvant therapy. *J Clin Oncol* 1998, **16**, 515–521.
13. Hayes DF, Trock B, Harris A. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? *Breast Cancer Res and Treat* 1998, **52**, 305–319.