The International Journal of Biological Markers, Vol. 18 no. 1, pp. 40-48

Biostatistics and tumor marker studies in breast cancer: Design, analysis and interpretation issues

E. Biganzoli[1], P. Boracchi[2], E. Marubini[2]

[1]Unità di Statistica Medica e Biometria, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan
[2]Istituto di Statistica Medica e Biometria, Università degli Studi di Milano, Milan - Italy

INTRODUCTION

The development of staging procedures for breast cancer aimed at the identification of distinct, homogeneous patient categories with different chances of benefiting from treatment has a long history. However, within the same TNM stage, patient outcome may still be highly variable, depending on differences in aggressiveness between tumors. For this reason, there is strong interest in the study of tumor markers to better discriminate subjects with different risks of disease recurrence and likelihood of responding to tailored adjuvant systemic treatments. Although several markers have been evaluated or are still under investigation, very few of them are considered clinically useful for predicting patient outcome or response to treatment (1, 2). Therefore, in spite of the continuous progress in the molecular biology of cancer, clinical decision-making still largely relies on pathological staging and grading procedures; in the case of breast cancer, there is a clear gap between the resources employed for basic and translational research on tumor markers and actual patient benefits and overall social gain.

The aim of this note is to provide an overview of the quantitative aspects related to the design and analysis of prognostic factor studies, and to show their critical role in determining the level of evidence of the results achieved. Several reviews have already been published to promote "good statistical practice" in tumor marker studies (3-6). Related issues will be reconsidered with emphasis on peculiar aspects of retrospective studies. To show the possible risks underlying the statistical evaluation of this kind of study, an example will be discussed in detail.

QUANTITATIVE ASPECTS IN PROGNOSTIC FACTOR ANALYSIS

Measurement scales

The issue of the reliability of tumor marker measurement has been widely addressed. Statistical methods play a key role in assay development and validation of the analytical performance of quantitative biochemical assays and qualitative pathological examination. In addition, quality control guidelines must be implemented for routine measurements, including internal quality control to confirm the stability of the precision and accuracy specifications that were determined in the validation phase, and external quality control, to assess the relative and absolute accuracy of assays in different laboratories by measuring the same reference samples (7).

A key point to be addressed is how the information provided by tumor markers in statistical models should be managed. The dichotomization of the measure into two groups (high vs low or positive vs negative) on the basis of a threshold value is still frequently applied. In principle, the cutoff point should be defined on the basis of sensible biological and/or clinical reasons. If these are not available, as is often the case, other criteria are adopted, a frequent one being the median value of marker distribution. However, there is no a priori reason to expect that exactly half of the patients, belonging to a group on the basis of the ordered values of the variable, have a better prognosis than the other half. In any case, before dichotomizing a continuous variable one should verify that the prognostic information provided by the variable in the original measurement scale is still adequately represented after dichotomization. It has been

0393-6155/040-09$04.50/0

shown that a substantial loss of information may occur when models with dichotomized variables are used in the absence of a reliable threshold in the underlying relationship between prognosis and the variable itself (8).

The quantitative scale provides the largest amount of information, while the categorization of measurements from a quantitative to a qualitative scale necessarily involves a loss of information that cannot be subsequently retrieved. These aspects must be taken into consideration when choosing the measurement scale by means of which data will be expressed.

Outcome measures

A quantity which is typically considered in the analysis of survival data is the event rate (λ). For a homogeneous population this rate is defined as the ratio between the number of observed events and the sum of the times during which the subjects were exposed to the risk of developing the event. This definition implies the need to consider the rate of appearance of the event as constant during the follow-up of the population under study. This assumption may be correct for some events, such as the occurrence of a contralateral tumor, but not for others, e.g. the appearance of distant metastases after breast cancer surgery. In the latter case it is necessary to divide the time of the study into intervals and calculate the risk rate for each interval. Considering intervals of decreasing size (i.e., tending to zero), an "instantaneous" estimate of the rate is obtained which is called *hazard*. The hazard is a useful measure for the clinician as it provides information about the relative speed of the occurrence of the event in time with reference to the number of studied subjects.

Regression models are adopted for the study of the relationship between the hazard and tumor/patient characteristics. The classical approach for the evaluation of the prognostic impact of tumor markers is the Cox model (9); it assumes a linear relationship between the logarithm of the hazard of an event and the covariates (prognostic factors).

Given the i-th subject with x_{ij} observed values for the j-th variable (j=1,2,$\cdots$,J), the Cox model can be written as:

$$\log \frac{\lambda(t, x_{ij})}{\lambda_0(t)} = \sum_j x_{ij}\beta_j$$

where λ (t,x_{ij}) is the hazard as a function of time and covariates, where $\lambda_0(t)$ is the event hazard for the category of subjects that have the value 0 for all the j covariates considered (reference category) and the regression coefficients β_j provide the estimate of the logarithm of the hazard ratio.

The standard formulation of the Cox model clearly implies that the ratio between the hazards is constant in time (*proportional hazard*); however, this assumption is not generally true and it should be carefully evaluated on a case-by-case basis. In the absence of proportional hazards, time-dependent effects can be considered in the Cox model. Since this model does not provide estimates of the hazard function, it is not a suitable choice to investigate disease dynamics or to make outcome predictions. Other flexible approaches may be adopted in an exploratory context, allowing assessment of the shape of the hazard function and the effects of continuous covariates (10, 11).

PROGNOSTIC AND PREDICTIVE FACTORS

According to the definition provided by Hayes et al (12), prognostic factors are associated with either the metastatic or the growth rate potential of the primary tumor. Predictive factors are associated with relative sensitivity and/or resistance to specific therapeutic agents. The same factor may be both prognostic and predictive. A pure prognostic factor discriminates between poor and favorable prognosis, independent of therapy. A pure predictive factor does not indicate the prognosis of untreated patients. However, in the presence of the specific treatment for which the factor predicts sensitivity or resistance, patient outcome will be different according to the values of the said factor.

Pure prognostic factors are associated with the natural history of the disease, for example the metastatic involvement of axillary lymph nodes. The estrogen receptor (ER) protein content of the primary tumor, on the other hand, has a predictive and a (minor) prognostic role. Patients with a high ER content in their tumors are much more likely to benefit from hormone therapy than those with low ER. Moreover, patients with low-ER tumors have a higher risk of relapse and death than those with high-ER tumors, thus indicating also the prognostic role of this marker.

Figure 1 reports different situations related to the role of the biomarker whose presence is protective. The upper panels (A, B, C) illustrate the situation where there is no treatment effect in subjects with negative marker levels, whereas in the lower panels (D, E, F) the treatment is effective also in subjects with negative marker levels. Panels A and D correspond to a pure prognostic effect of the marker, panels B and E to a pure predictive effect, and panels C and F to a combined prognostic/predictive effect. The situation of Figure 1C could represent the effect of estrogen receptor status and hormone therapy; in this situation the marker has a (slight) protective prognostic role (at least at early follow-up), whereas ER+ subjects are much more likely to respond to hormonal therapy, with a marked decrease in their risk.

The predictive effect is equivalent to the pharmacological concept of synergism (antagonism), namely, the combination of two factors (treatment and marker) produces a higher (lower) effect than the sum of the separate

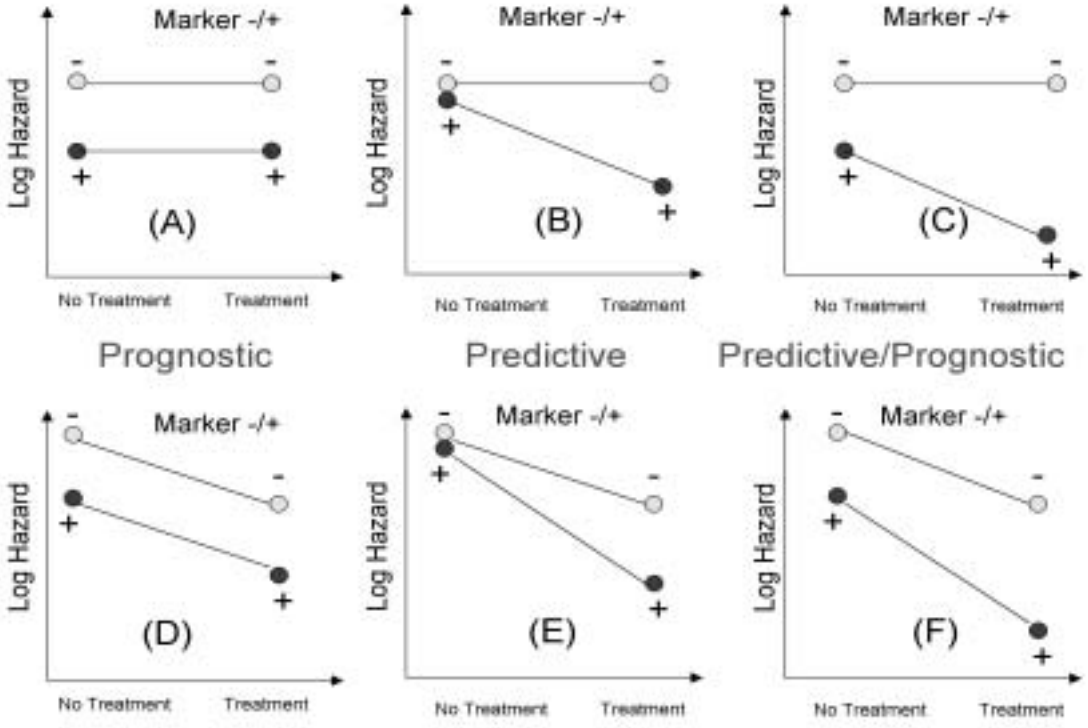

Fig. 1 - *Situations related to the role of a biomarker whose presence is protective. The upper panels (A, B, C) show the absence of any treatment effect in subjects with negative marker results, whereas in the lower panels (D, E, F) treatment is effective also in subjects with negative marker results. Panels A and D correspond to the purely prognostic effect of the marker, panels B and E to the purely predictive effect. Panels C and F correspond to the combined prognostic and predictive effects.*

effects of the two factors considered singly.

Prognostic and predictive effects can be statistically assessed by means of a Cox regression model, which in its basic form considers the treatment and the marker of interest:

$$\log \frac{\lambda(t, x_{ij})}{\lambda_0(t)} = \beta_M \chi_M + \beta_T \chi_T + \beta_{MT} \chi_M * \chi_T$$

including one term, β_M, for the effect of the marker, another, β_T, for the effect of treatment, and a third interaction term, β_{MT}, for the synergistic effect between treatment and marker. The interaction accounts for departures from the sum of the two separate effects of treatment and marker, which would be expected if the marker was not predictive (absence of synergism).

For each patient the marker status is coded as χ_M=0 if negative and χ_M=1 if positive; treatment is coded as χ_T=0 if absent and χ_T=1 if present, and the interaction is given by the product of the two variables χ_M*x_T. With this coding scheme there will be four possible combinations (shown in Table I), which characterize four patient groups according to marker and treatment status.

Figure 2 graphically reports the expected results of the Cox model in the different situations; these are described in detail in Table II.

More complex situations may also be considered, including markers with an unfavorable prognostic role and a favorable effect on therapy. In such cases the reversed situation is observed when treatment is applied. This is the so-called qualitative interaction effect, which is graphically represented in Figure 3. Such a situation

TABLE I

Group	χ_M	χ_T	χ_M*χ_T
1	0	0	0
2	0	1	0
3	1	0	0
4	1	1	1

TABLE II

Situation	Treatment		Prognostic marker		Predictive marker	
(A)	Ineffective	$\beta_T = 0$	YES	$\beta_M < 0$	NO	$\beta_{MT} = 0$
(B)	Ineffective	$\beta_T = 0$	NO	$\beta_M = 0$	YES	$\beta_{MT} < 0$
(C)	Ineffective	$\beta_T = 0$	YES	$\beta_M < 0$	YES	$\beta_{MT} < 0$
(D)	Effective	$\beta_T < 0$	YES	$\beta_M < 0$	NO	$\beta_{MT} = 0$
(E)	Effective	$\beta_T < 0$	NO	$\beta_M = 0$	YES	$\beta_{MT} < 0$
(F)	Effective	$\beta_T < 0$	YES	$\beta_M < 0$	YES	$\beta_{MT} < 0$

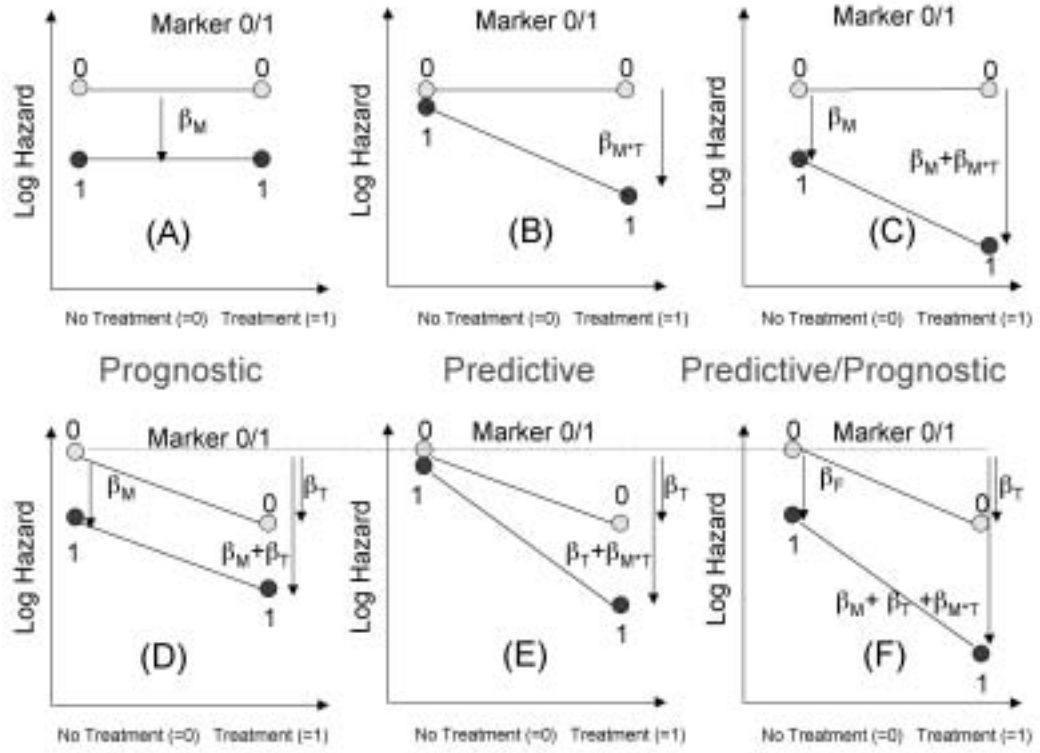

Fig. 2 - *Expected results of the Cox model in different prognostic/predictive situations. The panels correspond to those of Figure 1.*

could correspond, for example, to the relationship between c-*erb*B-2 and therapy with trastuzumab (Herceptin), which is targeted to this type of marker.

A single dichotomous marker was considered in the above examples. In the case of markers measured on a continuous scale the situation is much more complex because of the possible departures from linearity of the effects of the marker and of the interaction with treatment which need to be investigated. Nevertheless, a detailed study of continuous relationships may lead to an improvement of the decision criteria for a therapeutic strategy. For example, the optimal cutoff value for the levels of estrogen receptors that are required for efficacious hormonal therapy is still debated today. Very few studies have tackled this problem by considering the effect of ER measured on a continuous scale.

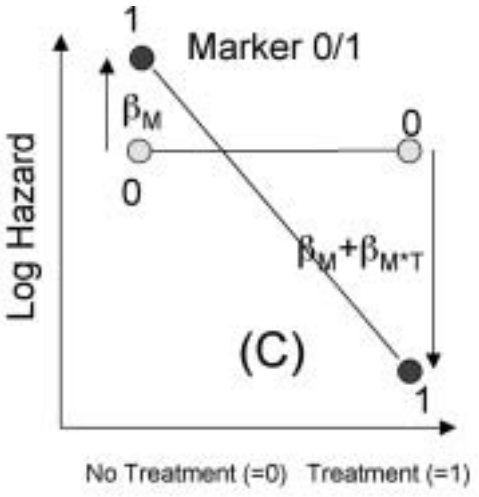

Fig. 3 - *Graphical representation of a qualitative interaction effect.*

STUDY DESIGN

A general problem of prognostic factor studies is their lack of standard design criteria. As a consequence, results may be contradictory, and although several studies regarding the same marker may be available, these have limited clinical and biological impact.

Criteria for the design of tumor marker studies have been proposed by Simon and Altman (3) and were subsequently revised by Altman and Lyman (4), who also proposed a classification of the different types of study into four classes:

1) **phase I -** exploratory (hypothesis generating) studies evaluating the association between new markers and recognized prognostic factors;
2) **phase II -** exploratory studies of the prognostic discrimination ability of the new marker or the response to therapy;
3) **phase III -** confirmatory study of prior hypotheses regarding prognosis or response to therapy; such studies are aimed at investigating the additional prognos-

tic contribution of a new factor to those already established;

4) studies for the development of outcome prediction models, combining the information of several prognostic variables to maximize the prognostic accuracy for groups and individual patients.

In our opinion the last category can be further divided into three subclasses, namely

4a) **pilot studies**, based on medium-sized case series (ratio between number of model terms and events around 1/10), evaluating the feasibility of development studies for outcome prediction models;

4b) **development studies**, based on large case series recruited on an institutional or regional basis (ratio between number of model terms and events higher than 1/10);

4c) **validation studies**, to assess the prognostic accuracy of outcome prediction models on new independent case series.

Prognostic factor studies can be prospective or retrospective; although the former are to be preferred, most are retrospective. The advantage of the latter is that medium- to long-term follow-up information may be already available. Disadvantages are related to possible selection biases due to missing data and the unavailability of stored specimens for marker determination. Retrospective studies may consider heterogeneous case series with respect to baseline clinical features and/or therapeutic strategies. For this reason, statistical analyses are not straightforward and their results may be difficult to interpret. A careful study design is needed for retrospective studies as well as for prospective ones.

Randomized controlled studies are reliable tools for the evaluation of therapeutic strategies. To improve the reliability of tumor marker studies, an approach similar to the design of randomized clinical trials should be adopted. This means that a study protocol should be prepared which clearly states the objectives of the study, including a rationale, and which explains the hypotheses to be tested. According to these points, patient inclusion and exclusion criteria must be specified for both prospective and retrospective studies, so that it is clear to which population of patients the results refer. Prognostic factors relevant for the study must be specified together with the methods adopted for their assessment (analytical methods for tumor markers). The outcome variable must be defined and the way in which it is assessed should be specified. The strategies for data analysis must be indicated and, finally, the criteria for interpretation and use of the results should be included.

Studies on pure prognostic factors should only be performed on series of patients who did not receive systemic therapy. If a marker is evaluated in patients who underwent a specific systemic treatment, it is not possible to disentangle pure prognostic effects from the response to therapy. Studies should therefore be limited to assessing whether the marker discriminates prognosis in patients treated with the specific therapy, without making any inference regarding the type of effect. When an interaction between treatment and marker exists, combining treated and untreated patients in analyses of potential prognostic factors makes it impossible to separate predictive from prognostic effects.

According to the above considerations, in retrospective evaluations the evidence of the predictive role of a marker with respect to a specific treatment should be assessed on the basis of comparison with a suitable control group in the context of a randomized clinical trial, whereas suitable designs have been proposed for prospective randomized studies (13).

Investigators might be interested in determining whether specific factors identify subsets of patients who do or do not benefit from an experimental treatment. However, statistical analyses of treatment effects in patient subsets identified by marker status are often unreliable. A suitable approach is to test directly for the presence of the interaction between treatment and the marker of interest. This is equivalent to a test for the homogeneity of treatment effects in the subgroups defined by marker status, which is more appropriate than separate tests for treatment effects in the same subgroups. However, it should be noted that clinical trials are seldom designed to test specific interactions of treatment with prognostic factors. This problem is particularly clear when marker measurements are done retrospectively on preserved specimens. The study may fail to highlight relevant interaction effects because they were not considered in the original design. Moreover, if several markers are considered in the same study, multiple interaction tests must be performed with caution because of the risk of spurious results which could occur by chance alone (see next section for statistical details). In this respect, suitable techniques have been proposed and applied to biological markers in breast cancer for the evaluation of response to alkylating agent-based adjuvant therapy (CMF) (14). Special attention should be paid to the conclusions of interaction analyses, which should be considered at an exploratory level unless the study was designed to test specific interactions between treatment and markers.

STATISTICAL INFERENCE

Testing hypotheses

In experimental studies the inductive approach is adopted to infer general properties of a population from the data of a sample of individuals belonging to the same population. Therefore, sampling variability affects the results; given the treatment and assuming that the marker is measured without error (this is not generally true), pa-

tients within the same treatment group with equal marker status have heterogeneous outcomes due to the inherent biological variability which cannot be controlled. Statistical hypothesis testing deals with such variability by associating a certain degree of uncertainty to the statements regarding the results of the study. The scheme of hypothesis testing (null hypothesis H_0: $\beta_j=0$) for the example regarding the effects of treatment and marker is reported in the following Table III:

TABLE III

"Truth" Decision	**True H_0: $\beta_j=0$**	**False H_0: $\beta_j\neq0$**
Do not reject H_0	Correct decision	Type II error (*False negative*) (Probability β)
Reject H_0	Type I error (*False positive*) (Probability α: statistical significance)	Correct decision (Probability 1-β: power of the test)

Many reported biomarker results in the literature are either falsely negative or falsely positive. The power of the test is the probability of detecting an effect when it really exists. Increasing the sample size increases power. Most studies of potential prognostic or predictive factors have low power, therefore a non-significant result may reflect a lack of statistical power rather than absence of the effect. Evaluation of a potential predictive biomarker is much more difficult than evaluation of a new therapy. The status of biomarkers cannot be randomized, and their imbalance must be taken into account. A test for an interaction between response to treatment and biomarker status usually requires many more events than a test for a treatment effect. Biomarker studies are often adjuncts to already completed randomized clinical trials that were designed to compare treatments. The subsets of patients with biomarker data are generally smaller than the randomized groups that received the treatments; therefore, given the hypothesis of a random unbiased selection of subjects with available measurements, at least the statistical power is reduced. As a result, most studies of potential predictive biomarkers may produce false-negative results.

However, many of the results about biomarkers in the literature are probably false-positive as a consequence of multiple tests of hypotheses. Performing subset analyses often leads to statistically significant results by chance alone. For example, this problem affects empirical attempts to identify a cutoff point to split biomarker measures into high-and low-risk categories on the basis of statistical significance alone.

AN EXAMPLE: c-*erb*B-2 AND RESPONSE TO ALKYLATING AGENT-BASED ADJUVANT THERAPY

Several studies addressed the question whether breast cancer patients treated with adjuvant CMF have a different benefit depending on the c-*erb*B-2 status of their tumors. In accordance with the considerations of the preceding paragraphs, only studies involving a control group without adjuvant treatment will be considered hereafter. We will discuss in particular the results of three studies involving retrospective determination of c-*erb*B-2 in patients recruited to randomized phase III clinical trials on chemotherapy, which potentially address the role of this marker as a predictor of the response to therapy. These studies were originally quoted in the review article of Yamauchi et al (15), who presented a case study of c-*erb*B-2 as a predictive factor in breast cancer.

In the first study of the ECOG/US Intergroup (16), 406 node-negative patients were randomized to receive either CMF with prednisone (CMFp, n=160 with c-*erb*B-2 determination by immunohistochemistry, IHC) or no systemic therapy (no ST, n=146 with c-*erb*B-2 determination). The authors performed separate log-rank statistical tests to evaluate the difference between the survival curves of patients in the two treatment arms according to c-*erb*B-2 status. Patients with c-*erb*B-2-negative tumors (231/306, 75.5%) showed a significant (p=0.003) improvement in disease-free survival (5-year DFS: c-*erb*B-2 negative, no ST: 58%; c-*erb*B-2 negative, CMFp: 80%) while patients with c-*erb*B-2-positive tumors (75/306, 24.5%) showed a non-significant improvement in DFS (5-year DFS: c-*erb*B-2 positive, no ST: 68%, c-*erb*B-2 positive, CMFp: 78%). It is worthy of note that in this study patients with c-*erb*B-2-positive tumors with no ST had better outcomes than c-*erb*B-2 negative patients, which was in agreement with the uncertain prognostic role of this tumor marker in node-negative breast cancer patients.

In the Guy's Hospital prospective randomized trials of adjuvant CMF therapy versus no therapy, specimens from 274 node-positive women (70% of 391 patients in the original trial) were evaluated for c-*erb*B-2 status by IHC (17). Of these patients 129 received 12 cycles of CMF and 145 received no ST. Both c-*erb*B-2-positive and negative patients gained benefit from adjuvant CMF. However, women whose tumors did not overexpress c-*erb*B-2 (191/274, 70%) had a significantly improved survival related to therapy (9-year overall survival: c-*erb*B-2 negative, no ST: 42%; c-*erb*B-2 negative, CMF: 65%; p=0.0014), whereas the difference between treated and untreated c-*erb*B-2-positive patients (83/274, 30%) did not reach conventional statistical significance (9-year OS: c-*erb*B-2 positive, no ST: 25%, c-*erb*B-2 positive, CMF: 42%; p=0.08).

In the prospective randomized trials of adjuvant CMF therapy versus no therapy conducted by the Istituto Nazionale Tumori of Milan, specimens from 337 node-

positive women (87% of 386 patients in the original trial) were evaluated for c-*erb*B-2 status and for other tumor markers by IHC (complete information was available on at least 324 cases) (14). Of these cases, 186 received 12 cycles of CMF and 151 received no ST. Statistical analysis directly investigated the interaction between adjuvant treatment and c-*erb*B-2 status. As interactions of treatment with several biomarkers were also considered in the Cox regression model, a Bayesian method was applied to avoid the application of multiple statistical tests. Results were reported as hazard ratios of CMF vs no ST. For relapse-free survival (RFS), the HR for c-*erb*B-2-positive patients (54/337; 16%) was 0.484 (0.284-0.827, 95% highest posterior density interval) and for c-*erb*B-2-negative patients (283/337; 84%) it was 0.641 (0.481-0.853, 95% HPD). Considering the cause-specific breast cancer survival (CSS), the HR for c-*erb*B-2-positive patients was 0.495 (0.287-0.853, 95% HPD) and for c-*erb*B-2 negative patients 0.730 (0.537-0.991, 95% HPD). These results support the evidence of the protective role of CMF on the RFS and CSS of the patients regardless of the c-*erb*B-2 status of their tumors.

Application of statistical inference concepts

To illustrate the application of statistical inference to the response to therapy, we will comment on the paper by Miles et al (17) regarding the Guy's Hospital trial. Figure 1 of this paper reports the overall survival curves according to treatment in the subgroups defined by c-*erb*B-2 status. The right part of the graph shows the value of the χ^2 test statistic for the difference between curves, the corresponding p-values, and the number of patients in each subgroup defined by treatment and c-*erb*B-2 status. A p-value of 0.0014 in the c-*erb*B-2-negative subgroup represents the probability of a false-positive result (Type 1 error), i.e., the probability of stating that the survival curves are different only by chance when they are really equal because the treatment has no effect. Since this probability is very low, it is reasonable to assume that the curves are different and the treatment was effective in improving survival. If we look at the c-*erb*B-2-positive subgroup, the χ^2 test statistic is much lower and a p-value of 0.08 is observed; therefore, with regard to this subgroup the probability of being wrong in declaring that the treatment was effective is higher because, if there was no effect of therapy in the same experimental conditions, the observed difference between survival curves would have occurred on average in 8% of times only by chance. This means that the null hypothesis of absence of treatment effect cannot be safely rejected. Now the question is whether a p-value of 0.08 in the c-*erb*B-2-positive subgroup, in the conditions of the Guy's Hospital trial, implies that the null hypothesis can be safely accepted, thus supporting the evidence of no effect of alkylating agent-based adjuvant therapy on c-*erb*B-2-positive patients. To verify the last point, we should consider the type II error probability, i.e., the risk of false negative results on the effect of therapy; before saying that there is evidence for the null hypothesis, it should be verified that the risk of stating no effect of therapy when it is actually effective is rather low. Unfortunately, this point is seldom considered either in subgroup analyses of clinical trial data or in critical reviews of the literature. In the example of the Guy's Hospital trial, we made an attempt at estimating the Type II error in the c-*erb*B-2-positive subgroup by reconstructing original survival times from the graph of Kaplan-Meier survival estimates. Following application of the approach proposed by Schoenfeld (18) for the difference between survival curves, we estimated a Type II error probability of about 0.4; that is, if alkylating agent-based adjuvant therapy was also effective for c-*erb*B-2-positive patients, on average, in the same experimental conditions, we would be able only about 60% of times to correctly reject the null hypothesis of no effect of therapy. This is the so-called power of the statistical test, which is typically fixed around 80-90% in clinical trials; in the subgroup analysis for c-*erb*B-2-positive patients of the Guy's Hospital trial it is unacceptably low. Such a low power is partly due to the distribution of patients for c-*erb*B-2 status, since positive tumors are much less frequent than negative ones.

A better approach would have been to test directly for the interaction between c-*erb*B-2 status and therapy. However, also this test would be likely to be underpowered, as the original trial was not specifically designed for testing the above interaction. The bottom line is that we cannot obtain conclusions from such analyses but only evidence for the hypotheses regarding treatment effect in patient subgroups. The critical aspect is the weight to be given to such evidence. The difference between statistical significance and practical relevance of a studied effect must be underlined. In the context of the example, p-values are used as a measure of evidence of the effect of interest: the smaller the p-value, the stronger is the evidence. It could be questioned whether p-values are a good measure of evidence (19), but we will not consider that aspect here. The use of p-values as a measure of evidence is different from the pure logic underlying statistical hypothesis testing; in the latter case the experiment is designed for rejecting or not rejecting the null hypothesis after fixing the α-level of type I error (typically 5%) and the power of the test (typically 80-90%); in such a context, actual p-values may not be reported because only the test decision is relevant ($p \leq \alpha$ or $p > \alpha$); in the exploratory analysis framework small p-values may occur because large case series are considered with enough information to highlight even small effects of little practical importance. By contrast, if little information is available, as in the case of the subgroup of c-*erb*B-2-positive patients of the Guy's Hospital trial, important effects may have high p-values only because the study is underpow-

ered. Therefore, the following aphorism is helpful: "Statistical significance does not imply practical importance and the lack of statistical significance (evidence) of an effect cannot be assumed as evidence of the lack of effect".

In addition to statistical significance, an estimate of the effect should be provided so as to allow assessment of its practical importance. According to the above considerations about the uncertainty of decisions regarding statistical hypotheses, also the estimate of the effect is uncertain, therefore a measure of such an imprecision should be provided. The use of confidence intervals has the advantage of providing a range of values in which the true effect lies, with a fixed confidence (probability of being right), in making such a statement. In the case of the subgroup of c-*erb*B-2-positive patients of the Guy's Hospital trial, we estimated a treatment effect of 0.64 measured as hazard ratio under the proportional hazards assumption, with 95% confidence intervals: 0.38-1.07. As regards the subgroup of c-*erb*B-2-negative patients, from the estimated survival curves we calculated a hazard ratio of about 0.45, which is contained in the above confidence interval; therefore, the evidence of a different effect of alkylating agent-based adjuvant therapy according to c-*erb*B-2 status appears difficult to uphold on the basis of the Guy's Hospital trial results. Similar considerations about the results of a subgroup analysis concerning the same problem have been provided by Simon and Altman (3) with regard to the study by Gusterson et al (20).

*Treatment guidelines on alkylating agent-based adjuvant therapy according to c-erb*B-*2 status*

The College of American Pathologists Consensus Statement 1999 (21) reported three different categories of prognostic factors according to specific rules: category I factors are proven to be of prognostic importance and useful in clinical patient management; category II factors have been extensively studied biologically and clinically, but their importance remains to be validated in statistically robust studies; category III groups all factors not sufficiently studied to demonstrate their prognostic value.

The tumor marker c-*erb*B-2 belongs to category II, given the issues discussed above. c-*erb*B-2 is considered to be associated with lower responsiveness to methotrexate-based treatment regimens; however, several questions have been raised regarding the standardization of c-*erb*B-2 analysis.

The 2000 Update of Recommendations for the Use of Tumor Markers in Breast and Colorectal Cancer: Clinical Practice Guidelines of the American Society of Clinical Oncology (1) in the paragraph "Response to Cyclophosphamide/Methotrexate/Fluorouracil or Nonanthracycline-Based Adjuvant Chemotherapy", after reviewing the literature, states that, "The question of whether c-*erb*B-2 overexpression affects the relative benefit of adjuvant CMF chemotherapy remains open, and the Update Committee cannot make a definitive practice recommendation at present."

In their review article Yamauchi et al (15) referred to the results of the Guy's Hospital study (17) with the following statement: "Two conclusions can be reached from this study. The results support the role of c-*erb*B-2 as a negative prognostic factor, because untreated c-*erb*B-2-positive patients had worse DFS and OS than c-*erb*B-2-negative patients independent of therapy. Moreover, although within each expression group one can observe benefit from adjuvant CMF treatment versus no treatment, the magnitude of benefit seems larger for those whose tumors are c-*erb*B-2 negative."

Referring to the Milan study (14), Yamauchi et al (15) report: "In contrast, c-*erb*B-2 analysis of patients that were included in the Milan CMF versus observation trial has produced contradictory results. (···) Indeed, if anything, c-*erb*B-2-positive patients were more likely to benefit."

CONCLUSIONS

The reader is invited to re-examine the statements of the above guidelines and reviews according to the few basic statistical principles provided in this paper, in order to decide whether the claimed evidence regarding c-*erb*B-2 and response to alkylating agent-based adjuvant therapy has to be reconsidered.

An incontrovertible aspect is the lack of conclusive evidence. If the statistical principles considered in this paper were applied at least to the review of the literature, in the authors' view, the evidence of a different effect of alkylating agent-based adjuvant therapy according to c-*erb*B-2 status would have received less support and the results of the Milan study (14) would not have been considered contradictory.

Overall, balanced conclusions of such studies and their reviews should stress the need for suitably designed trials to obtain conclusive evidence regarding treatment strategies according to marker status. Such considerations, based on a relevant problem in adjuvant therapy for breast cancer, can be extended to analyses of the prognostic and predictive role of other biomarkers. The issues discussed in the present paper appear highly relevant when the impact of inappropriate evaluation of biomarker data is considered in terms of suboptimal therapies and research costs.

ACKNOWLEDGEMENTS

The authors thank Dr. S. Alberti for his help in editing the manuscript.

This work was partially supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC).

Address for correspondence:
Dr. Elia Biganzoli
Unità di Statistica Medica e Biometria
Istituto Nazionale per lo Studio e la Cura dei Tumori
Via Venezian, 1
20133 Milano, Italy
email: biganzoli@istitutotumori.mi.it

REFERENCES

1. Bast RC Jr, Ravdin P, Hayes DF, et al. 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. J Clin Oncol 2001; 19: 1865-78.
2. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Senn H-J. Meeting highlights: International Consensus Panel on the treatment of primary breast cancer. J Clin Oncol 2001; 19: 3817-27.
3. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994; 69: 979-85.
4. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 1998; 52: 289-303.
5. Pajak TF, Clark GM, Sargent DJ, McShane LM, Hammond ME. Statistical issues in tumor marker studies. Arch Pathol Lab Med 2000; 124: 1011-5.
6. Clark GM. Interpreting and integrating risk factors for patients with primary breast cancer. J Natl Cancer Inst Monogr 2001; 17-21.
7. Paradiso A, Volpe S, Iacobacci A, et al. Quality control for biomarker determination in oncology: the experience of the Italian Network for Quality Assessment of Tumor Biomarkers (INQAT). Int J Biol Markers 2002; 17: 201-14.
8. Biganzoli E, Boracchi P, Daidone MG, Gion M, Marubini E. Flexible modelling in survival analysis. Structuring biological complexity from the information provided by tumor markers. Int J Biol Markers 1998; 13: 107-23.
9. Cox D. Regression models and life tables (with discussion). J R Stat Soc B 1972; 34: 187-200.
10. Boracchi P, Biganzoli E, Marubini E. Joint modelling of cause specific hazard functions with cubic splines: an application to a large series of breast cancer patients. Comp Statistics Data Analysis 2002 (in press).
11. Biganzoli E, Boracchi P, Marubini E. A general framework for neural network models on censored survival data. Neural Networks 2002; 15: 49-58.
12. Hayes DF, Trock B, Harris AL. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? Breast Cancer Res Treat 1998; 52: 305-19.
13. Boracchi P, Biganzoli E. Markers of prognosis and response to treatment: ready for clinical use in oncology? A biostatistician's viewpoint. Int J Biol Markers 2003 (in press).
14. Menard S, Valagussa P, Pilotti S, et al. Response to cyclophosphamide, methotrexate, and fluorouracil in lymph node-positive breast cancer according to HER2 overexpression and other tumor biologic variables. J Clin Oncol 2001; 19: 329-35.
15. Yamauchi H, Stearns V, Hayes DF. When is a tumor marker ready for prime time? A case study of c-erbB-2 as a predictive factor in breast cancer. J Clin Oncol 2001; 19: 2334-56.
16. Allred DC, Clark GM, Tandon AK, et al. HER-2/neu in node-negative breast cancer: prognostic significance of overexpression influenced by the presence of in situ carcinoma. J Clin Oncol 1992; 10: 599-605.
17. Miles DW, Harris WH, Gillett CE, Smith P, Barnes DM. Effect of c-erbB(2) and estrogen receptor status on survival of women with primary breast cancer treated with adjuvant cyclophosphamide/methotrexate/fluorouracil. Int J Cancer 1999; 84: 354-9.
18. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981; 68: 316-9.
19. Marubini E. Evidenza scientifica e statistica in la salute tra informazione e prova scientifica. In: Gallo C, Signoriello G, eds. SISMEC Conference Proceedings, 2001; 11-8.
20. Gusterson BA, Gelber RD, Goldhirsch A, et al. Prognostic importance of c-erbB-2 expression in breast cancer. International (Ludwig) Breast Cancer Study Group. J Clin Oncol 1992; 10: 1049-56.
21. Fitzgibbons PL, Page DL, Weaver D, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 2000; 124: 966-78.