9

Statistical Considerations in Assessing Molecular Markers for Cancer Prognosis and Treatment Efficacy

James Dignam, John Bryant, and Soonmyung Paik

1. Introduction

The development and growth of molecular biologic technology is leading to a new appreciation of inherent heterogeneity in cancer. While long appreciated as morphologically diverse entities, malignancies have increasingly been explored for molecular characteristics indicative of cellular regulation and growth, ability to adapt to and change local environments, and susceptibility to potentially therapeutic agents. These pursuits have led to important advances in the understanding of cancer biology, and in selected instances, have led to the development and use of treatments designed to act on molecular targets.

Depending on the technology used to obtain the data, the evaluation of molecular disease characteristic markers in relation to outcomes may involve novel statistical analysis problems, as well as familiar design and analysis issues. The recent introduction of DNA microarray technology, in which dozens or even hundreds of molecular characteristics of a tumor can be quantified and compared to normal tissue or to other tumors, is a relevant example. Researchers are interested in which of these molecular markers may be indicative of poorer outcomes or response to specific therapies. An appropriate evaluation of this extremely large volume of data challenges the limits of current statistical methodology.

In this chapter, we examine a current research question in breast cancer biology as an illustrative example to circumscribe methods for the analysis of new molecular markers in relation to clinical outcome data. Specifically, the clinical utility of a molecular characteristic of breast cancer tumors is evaluated, using archived tumor samples combined with clinical follow-up information

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

collected from a randomized clinical trial. This marker, the overexpression of the erbB-2 (also referred to as HER2/neu) protein on the cell surfaces of breast tumors, can potentially be used to select which chemotherapy drugs are liable to be of most benefit, and also has led to the development of new treatment agents designed to target the growth factor receptor encoded by the *erbB-2* oncogene.

2. Prognostic and Predictive Markers In Cancer

To better anticipate outcomes and tailor treatment for individuals with cancer, factors potentially indicative of prognosis have been investigated and employed in clinical decision-making. The extent of disease development and spread at time of diagnosis, usually a composite of features collectively referred to as the stage, is an important prognostic factor in all cancers. Related characteristics, such as size of the tumor, as well as the predominant tumor cell type and other pathologic features, are also well-recognized indicators of prognosis. Additional specific tumor cell characteristics, including the expression of receptors and protein complexes on tumor cell surfaces and the presence of genes in mutated form, may be associated with poor prognosis and/or poor response to treatment.

On this latter note, an important concept popularized in cancer studies but possibly unknown to statisticians (or known by another name) relates to factors that predict differential response to therapy in absence of or in addition to any relationship to prognosis in general. The term *prognostic factor* is generally reserved for those factors that identify patients at increased risk of relapse or death, as in the case of stage mentioned above. Factors that preferentially identify patients who respond to a given treatment are referred to as *predictive factors*. For example, tumors with certain pathologic characteristics may appear insensitive to chemotherapy. A characteristic can be both prognostic and predictive, an example being estrogen receptors found on the surface of breast tumor cells. The absence of such receptors is indicative of loss of cellular regulation and generally more profound pathologic aberrations leading to poorer outcomes. Furthermore, those patients with estrogen-receptor-bearing tumors have been found to be amenable to treatment with tamoxifen, an estrogen-like compound that blocks receptors and inhibits cell growth. Thus, estrogen receptors are both a prognostic marker and are predictive of treatment response with a specific, targeted agent. In this discussion, we will generally refer to markers under evaluation as prognostic markers, with the understanding that such markers will be evaluated for any relationship to treatment efficacy as well. Factors that might segregate patients who do not require chemotherapy after surgery from among those with early stage breast cancer are of particular interest, as there remains considerable debate regarding the worth of such treatment among so-called ''good risk'' patients ***(1–3)***.

Clinical utility of a marker is generally defined as the circumstance whereby knowledge of the marker value can prompt clinical action, including increased

diagnostic vigilance or specific treatment administration, which may benefit the patient. Because there has been a proliferation of potential markers in cancer, and yet little progress in achieving clinical utility for most, efforts have been introduced to define guidelines and criteria for new marker evaluation. The College of American Pathologists (CAP) has recently defined a three-level category ranking system for prognostic factors ***(4)***. Expert panels comprised of pathologists, cancer biologists, clinicians, statisticians, and others have been convened periodically to deliberate on the substance and quality of evidence for prognostic factors in cancer. CAP category I factors have proven value established in several studies, preferably prospective trials where marker evaluation was a study objective. Category II factors are those with evidence of utility that require further study and verification. Category III factors are generally new markers with limited data available thus far. These include anecdotal and small data observations, usually accompanied by a substantive underlying biological motivation.

The concept of level of evidence (LOE) has been established for the evaluation of data concerning therapeutic interventions (described at the website http://cancernet.nci.nih.gov/clinpdq). It has been proposed that a similar scheme be applied to prognostic marker studies, so that physicians and other scientists can more uniformly and objectively evaluate the literature and better develop a research agenda to address outstanding questions. **Table 1** shows the LOE evaluation criteria suggested by Hayes and colleagues as part of a comprehensive system to evaluate markers for clinical utility ***(5)***. Their TMUGS (Tumor Marker Utility Grading System) was developed in response to the somewhat haphazard manner in which marker information has developed over time, contributing to the relatively small improvement in prospective clinical evaluation of cancer patients. The LOE scale is applied to available clinical studies and, using this information together with an assessment of the assay methods, a semiquantitative score is derived reflecting to what extent evaluation of patients for the marker should become part of routine clinical decision-making.

It should be noted that, while a study satisfying the CAP category I or LOE I criteria would be ideal for unequivocally establishing the role of a new marker, such studies are unlikely to be conducted. The financial resources available for studies focused on markers rather than potentially therapeutic interventions are limited, and there are ethical implications of increasing sample size for therapeutic clinical trials to accommodate adequately powered ancillary studies of prognostic markers. Despite these barriers to the conduct of optimally designed marker studies, there is substantial opportunity for improvement of such studies within practical limitations, as discussed in the remainder of this chapter.

Table 1
Levels of Evidence for Grading Clinical Utility of Tumor Markers

Level	Type of evidence
I	Evidence from a single, high-powered, prospective, controlled study (with therapy and follow-up dictated by protocol) specifically designed to test marker or evidence from meta-analysis and/or overview of level II or level III studies. Ideally, study is a prospective, controlled randomized trial in which diagnostic and/or therapeutic clinical decisions in one arm are determined at least in part on the basis of marker results, and diagnostic and/or therapeutic clinical decisions in the control arm are made independently of marker results. However, study design may also include prospective but not randomized trials with marker data and clinical outcomes as the primary objective.
II	Evidence from a study in which marker data are determined in relationship to prospective trial that is performed to test therapeutic hypothesis but not specifically designed to test marker utility (i.e., marker study is secondary objective of protocol). However, specimen collection for marker study and statistical analysis are prospectively determined in protocol as secondary objectives.
III	Evidence from large but retrospective studies from which variable numbers of samples are available or selected. Therapeutic aspects and follow-up of patient population may or may not have been prospectively dictated. Statistical analysis for tumor marker was not dictated prospectively at time of therapeutic trial design.
IV	Evidence from small retrospective studies that do not have prospectively dictated therapy, follow-up, specimen selection, or statistical analysis. Study design may use matched case controls, etc.
V	Evidence from small pilot studies designed to determine or estimate distribution of marker levels in sample populations. Study designs may include "correlation" with other known or investigational markers or outcome but is not designed to determine clinical utility.

Adapted from the Tumor Marker Utility Grading System of Hayes et al. ***(5)***, with permission from Oxford University Press.

3. Statistical Issues in Prognostic Marker Studies

Statistical issues to be considered in prognostic marker studies are numerous. First, the method in which the marker is acquired may involve laboratory assays and procedures in which reproducibility and validity are concerns. Furthermore, there may be competing laboratory evaluation methods with different scoring systems for a given marker, and various discrete cut-points used for classification of assay results into positive or negative findings. An appropriate evaluation of a new marker must take into account existing prognostic factors, as disease characteristics are often correlated, and a new marker

may add little additional information over established factors (which may be easier and more economical to obtain). Thus, modeling with multiple covariates is required, and statistical power for these models may be inadequate, particularly when evaluating whether there exists any differential treatment response associated with marker values, represented by interaction terms in the model. Additional problems include multiplicity issues associated with examining multiple cut-points for a marker and examining multiple related outcome measures. Several excellent summaries of statistical problems in prognostic factor studies have appeared in recent years, and in this chapter we reiterate much of this work ***(6–8)***. Specific issues related to evaluation of erbB-2 in relation to breast cancer are discussed throughout **Subheading 4**.

3.1. Assay Evaluation

Any laboratory procedure is subject to measurement error, and modern molecular biology techniques in particular may involve complex processes that must be carefully controlled. Validity of results from such assays must be established through standard sensitivity and specificity evaluation, provided that a ''gold standard'' evaluation method and result are available. For new markers and techniques, such a standard may not be available, and expert consensus may be required to standardize and score results. (*See* Chapter 5 for a discussion of many of these issues.) In addition, inter-laboratory variability may need to be accounted for, as many studies in cancer involve the enrollment of patients from multiple institutions where laboratory quality and practice may differ. Finally, most tumor marker studies are conducted retrospectively on archived materials that may be sub-optimal, and it is important to address the validity of findings from such studies in relation to the types of samples that might be used in prospective evaluation of patients in clinical practice.

3.2. Scoring and Classification of Marker Results

The choice of cut-points for discrete classification of assay results is often not well motivated. When multiple classification schemes are investigated, and the grouping that produces the largest difference in outcome subsequently selected, the result can be a serious inflation of the apparent prognostic value of the marker, and other studies may then fail to reproduce the observation ***(9,10)***. Consequences of evaluating numerous cut-points were illustrated in a study by Hilsenbeck and Clark ***(11)***. In their study, simulations were conducted whereby multiple cut-points were applied to a continuous null marker (e.g., with no prognostic significance) to create two groups that were then compared in relation to clinical outcomes. Type I error rates increased from the expected 5% to 20–25% and higher when 5–10 candidate cut-points were tested and the maximum test statistic obtained was taken as the overall result of the marker

evaluation. The authors also provided a review and comparison of methods for adjustment of *p*-values obtained from testing of prognostic markers. Cut-points might be avoided altogether by using continuous marker values, the functional form of which might be obtained by various exploratory methods such as splines ***(12–14)***.

3.3. Statistical Power and Modeling

Statistical power is often inadequate in prognostic factor studies, and because most such studies are retrospective and observational in nature, the problem is further exacerbated by lack of a randomization mechanism, missing or misclassified covariates, and other problems. Several authors have commented on sample size requirements for adequate detection of main and interaction effects in prognostic factor studies. For simplicity of the discussion, we assume here that the marker can be partitioned into a dichotomy. The most common effect measure in prognostic factor studies with survival or related time-to-event endpoints is the risk ratio, which is usually computed from the Cox proportional hazards model ***(15)***. No definitive rule exists for effect size, but generally a marker that imparts a risk of twofold or greater would be considered to have clinically consequential potential. For the multiplicative relative risk scale, markers that impart small risks are not likely to be found statistically significant in small samples. The frequency distribution of the marker values will also influence statistical power, and in general these frequencies cannot be manipulated but are subject to the observed prevalence of the marker. In most cases, prevalence of unfavorable values for the marker will not be near the optimal value of 50%. Schoenfeld derived the required sample size for the Cox proportional hazards model, obtaining the same formula as that for two-sample log-rank test comparison under the proportional hazards assumption ***(16)***. For a given two-sided significance level α, power $1-\beta$, and risk ratio *(RR)* of interest, the total number of failures required is

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1\beta}\right)^2}{\left(1n(RR)\right)^2 \omega(1-\omega)},$$

where $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are $100 \times (1 - \alpha/2)\%$ and $100 \times (1 - \beta)\%$ standard normal deviates, respectively**,** and ω is the proportion of patients with the marker value of interest.

Schmoor and colleagues have extended Schoenfeld's results to account for correlation between the marker of interest and some other covariate, as analysis of new prognostic factors necessitates the consideration of known prognostic markers ***(17)***. Their derivation results in a straightforward modification of

the above equation that incorporates a "variance inflation factor" to account for correlation between model covariate X_1 and another covariate (or composite of other covariates) X_2:

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1\beta}\right)^2}{\left(\ln(RR)\right)^2 \omega(1-\omega)} \cdot \left(\frac{1}{1-\rho^2}\right),$$

where ρ is the correlation between X_1 and X_2.

For interaction effects, the situation is more challenging. Petersen and George ***(18)*** addressed sample size requirements for study designs of interaction effects in $2 \times K$ factorial experiments, where there are two treatments and a marker takes on $k = 2, 3,.., K$ values. Again, a modification of the usual sample size formula for hazard ratios is obtained. For the case of two treatment groups ($i = 1, 2$) and a two level prognostic marker ($j = 1, 2$), we define $\Delta_1 = \lambda_{11}/\lambda_{21}$, the treatment hazard ratio for level 1 of the marker and $\Delta_2 = \lambda_{12}/\lambda_{22}$, the treatment hazard ratio for level 2 of the marker. We wish to test H_o: $\Delta_1/\Delta_2 = 1.0$ using a two-sided α level test with power $1 - \beta$ against a specific interaction effect $\Delta_1/\Delta_2 = \theta \neq 1.0$. Under a proportional hazards assumption, the estimator $\ln(\Delta_1/\Delta_2)$ has variance approximately equal to $\sum_{ij} 1/n_{ij}$, where n_{ij} is the number of failures observed in treatment i and marker level j. It follows that the number of failures needed to achieve power $1 - \beta$ must approximately satisfy

$$\left(\sum_{ij} 1/n_{ij}\right)^{-1} = \frac{\left(Z_{1-\alpha/2} + Z_{1\beta}\right)^2}{\left(\ln(\Delta_1/\Delta_2)\right)^2}.$$

Using the fact that the harmonic mean of the n_{ij}'s is less than or equal to the arithmetic mean, this equation shows that the total number of failures required to detect a treatment by marker interaction with power of $1 - \beta$ is at least four times greater than the number of failures needed to detect a similarly sized treatment hazard ratio within a population that is homogeneous with respect to the prognostic marker. In designed experiments where treatment allocation could be balanced within strata of marker values via prospective sampling (so that $n_{1j} \approx n_{2j}$ for $j = 1, 2$), then the "four times greater" rule holds well; in cases of unequal frequencies of n_{ij}, sample size requirements will be even larger. Schmoor and colleagues also addressed interaction effects taking other covariates into consideration for the case of exponential failure times, which provides an approximate solution for the more general case ***(17)***.

In addition to statistical power, there are numerous other statistical considerations in prognostic marker studies specifically related to modeling. These

include verification of the correct model form, variable selection methods (e.g., stepwise regression and others), the aforementioned issues concerning definitions of discrete covariates, and model validation on independent data. A detailed discussion of concerns related to modeling can be found in Simon and Altman ***(6)*** and George ***(7)***. Klinger and colleagues discuss some alternatives to typical survival analysis modeling methods, such as regression trees, in the context of oncology research ***(19)***.

4. Case Study: *erbB-2* and Breast Cancer Treatment Response

4.1. Background: erbB-2 and Breast Cancer

The *erbB-2* oncogene (also known as c-*erbB-2* and HER2/*neu*), which encodes a specific transmembrane growth factor receptor of the tyrosine kinase family, was found to be amplified in a human breast carcinoma cell line by King and colleagues in 1985 ***(20)***. Subsequently, Slamon and colleagues reported that amplification of the *erbB-2* gene was present in 20–30% of breast cancers and was associated with shorter survival and disease-free survival time ***(21,22)***. It was conjectured that the basis for this association was a greater cell proliferation rate in tumors with erbB-2 amplification. These and subsequent analyses showed that erbB-2 (either amplification of the gene or overexpression of its protein product) was prognostic among both patients with tumors that had spread to the axillary lymph nodes (node-positive patients) and among patients with tumors confined to the breast (node-negative patients) ***(23–25)***. Other reports, however, did not confirm the relationship, or did not show a strong independent prognostic value for erbB-2, and there has been controversy in establishing the role of the marker as a clinically useful prognostic factor ***(26–29)***. Some authors have related this controversy directly to issues concerning laboratory evaluation of the marker ***(23,30–32)***. Nevertheless, the weight of evidence currently suggests that overexpression of erbB-2 does impart a less favorable prognosis. A recent meta-analysis of approx 35 studies appearing between 1996 and 1999 found erbB-2 to be a moderate but not particularly strong risk factor for breast cancer recurrence and death ***(33)***.

Early studies of erbB-2 suggested that it was not only associated with poor prognosis but also with a differential benefit depending on the chemotherapy drug or regimen administrated. Several studies suggested that tumors with overexpression did not respond as well to cyclosphosphamide, methotrexate, and fluorouracil (CMF, a commonly used chemotherapy regimen) as erbB-2 negative tumors ***(26,27,34)***, while others did not confirm this finding ***(35,36)***. Other studies suggested that erbB-2 overexpressing tumors were less sensitive to tamoxifen ***(37)***, again an observation not confirmed by others ***(38,39)***.

Overexpression of erbB-2 was more convincingly correlated with response to regimens containing doxorubicin (commercially, Adriamycin, a member of

a class of agents known as anthracyclines) in a series of studies appearing in the middle to late 1990s. Muss and colleagues first reported that a more intensive dose of cyclophosphamide, doxorubicin, and fluorouracil was of greater benefit among erbB-2 positive patients than among those with tumors that did not overexpress erbB-2 ***(40)***. A subsequent analysis of the same patient cohort and additional data also supported this conclusion ***(41)***. Independent reports, one of which is discussed in **Subheading 4.2.**, have also supported an association between overexpression and response to doxorubicin, and thus suggest that one might use the marker in clinical practice to choose treatment, as least for this agent ***(42,43)***. Additional investigations have explored whether those with overexpression would preferentially benefit from taxanes, but little reliable information is available thus far. Finally, a targeted agent for the erbB-2 receptor, trastuzumab (Herceptin, commercially), has appeared to show preferential efficacy among tumor cells overexpressing erbB-2 in preclinical studies ***(44)***. Thus far, efficacy trials of trastuzumab in humans have been conducted exclusively among erbB-2 positive patients. Assuming the mechanism of action is correct, it is plausible that little or no benefit would be realized for this agent among those whose tumors are erbB-2 negative.

4.2. The National Surgical Adjuvant Breast and Bowel Project B-11 Trial

The National Surgical Adjuvant Breast and Bowel Project (NSABP) is a federally funded multicenter cooperative clinical trials group that has carried out studies addressing the treatment and prevention of breast and colorectal cancers. A spectrum of modalities has been investigated, including surgical procedures, radiotherapy, chemotherapy, hormonal therapy, and biologic agents. In parallel with this effort, pathologic materials are collected and analyzed to investigate ancillary questions in the natural history and treatment of these cancers. Pathology materials are evaluated concurrently with conduct of the studies, and are also archived for future use.

In an earlier NSABP study, erbB-2 protein overexpression was found to be associated with poorer survival prognosis and other unfavorable pathologic features among patients with either node-negative or node-positive breast cancer ***(45)***. Subsequently, the potential for differential response to therapy according to erbB-2 status was investigated in NSABP protocol B-11, a randomized clinical trial evaluating the addition of doxorubicin to a two-drug chemotherapy regimen of L-phenylalanine mustard and fluorouracil (denoted PF) ***(43,46)***. Although PF has been superseded as a treatment option for breast cancer and other trials were available for erbB-2 evaluation, because the B-11 regimens differed only by the addition of doxorubicin, the trial was selected for evaluation first as a ‘‘proof of principle’’ study to address the potential

erbB-2–doxorubicin interaction. What follows is a detailed description of the analysis.

In protocol B-11, women with lymph node positive operable breast cancer were treated by either radical or modified radical mastectomy and randomized to receive either (1) PF or (2) PF and doxorubicin (PAF). Between June 1981 and September 1984, 707 patients were randomized, of whom 682 met study eligibility requirements. Further details of the study design and primary findings have been published previously ***(46)***. Endpoints for evaluation of erbB-2 in relation to response to doxorubicin were the same as those for the primary analysis of B-11. Disease-free survival (DFS) time was defined as time from surgery until breast cancer recurrence at any local, regional, or distant anatomic site, new primary cancer of any site, or death prior to these events. Survival time was defined as time until death from any cause. Two additional secondary endpoints, distant disease-free survival (DDFS) and recurrence-free survival (RFS), were addressed in the study but are not presented here.

4.2.1. Evaluation of erbB-2 *in NSABP Protocol B-11 Tumor Samples*

While about 200 patients had paraffin-embedded tumor blocks, > 90% (638 patients) had archived precut unstained tumor sections or hematoxylin–eosin (H&E) stained sections prepared as slides, and, consequently, these materials were used to perform evaluation for erbB-2. Such material is amenable to immunohistochemical (IHC) analysis to determine erbB-2 protein overexpression. IHC staining was performed using a cocktail of two antibodies (described in detail in ***[43]***) using both the unstained materials and stained slides. The determination of whether there was overexpression was based on a simple dichotomy, whereby the reaction was scored as positive if any cells showed definitive staining. Two individuals rated the slides together while blinded as to treatment assignment or outcome of the patient.

In this analysis, one immediate concern was whether the unstained and stained slides could be similarly stained and interpreted for erbB-2 and thus a simple sensitivity and specificity analysis was conducted. A comparison of staining sensitivity was performed whereby the assays of 51 cases were replicated using both stained and unstained sections available on the same patients. A simple cross-tabulation of positive and negative findings according to the two methods indicated 98% agreement (50 of 51 cases). Another quality assessment of materials involved a comparison of freshly cut sections from the paraffin blocks and previously cut and prepared slides. Sixty cases for which paraffin blocks and slides were available were assessed, and a 12% false-negative rate (25 were positive in fresh section, 22 were positive in slides) was observed. Thus, an analysis based on slide materials may be biased toward an attenuation of the effect of erbB-2 positivity, in that erbB-2 positive cases may be classified as negative.

Scoring methods for IHC and other assays of erbB-2 have been the subject of considerable controversy ***(47)***. For the IHC analysis results in the NSABP study, the percentage of positive staining for each patient's specimen was computed. The distribution was highly bimodal at 0% and 100%, suggesting that the dichotomous classification was the best approach with the available material. A further rationale for choosing the dichotomous rating system was that, in unstained and H&E stained slides rather than fresh material, it was deemed difficult to ascribe meaning to the quantitative percentage of cells staining, as has been done by other investigators, because the result could be largely an artifact of the laboratory procedure. Questions of assay reliability for IHC methods have led some to suggest that fluorescence *in situ* hybridization (FISH) analysis, which measures gene amplification (copy number) rather than protein expression, would be preferable ***(24,30,48)***.

4.2.2. Relationship of erbB-2 *to Other Patient and Tumor Characteristics*

Because negative and positive prognostic factors are often interrelated, the joint distribution of erbB-2 overexpression and other patient and tumor characteristics was examined. About 37.5% of patients exhibited erbB-2 positive tumors. Examining the cross-classification of factors singly with erbB-2, it was found that a higher number of positive nodes, larger tumor size, and estrogen receptor negative tumors were associated with erbB-2 overexpression. To take factors into account jointly, a logistic regression model relating erbB-2 to all covariates was employed, yielding similar results.

4.2.3. erbB-2 *as a Predictor of Response to Doxorubicin*

The explicit hypothesis of this investigation was that the benefit of doxorubicin would be largely confined to those patients with erbB-2 overexpression, that is, outcomes would differ in favor of PAF among patients with overexpression, while among those without overexpression, outcomes for PF and PAF would be similar. Accordingly, comparisons of treatment outcomes were conducted separately for the cohorts of erbB-2 negative (n =399) and erbB-2 positive (n = 239) patients. Kaplan–Meier estimates of disease-free survival (DFS) and survival are shown in **Fig. 1**. For each erbB-2 cohort, a PAF/PF relative risk *(RR)* estimate and corresponding significance test for the null hypothesis $RR = 1.0$ were obtained by the Cox proportional hazards model containing other relevant prognostic covariates (patient age at surgery, clinical tumor size, lymph node status, and estrogen receptor status). Results suggested that the benefit of doxorubicin (e.g., the PAF treatment arm) was evident only for those patients overexpressing erbB-2. This was confirmed by a formal test of differential benefit for doxorubicin according to erbB-2 status by combining all patients and testing an interaction term in the proportional hazards model **(Fig. 2)**. The resulting interaction tests for the various endpoints were statisti-

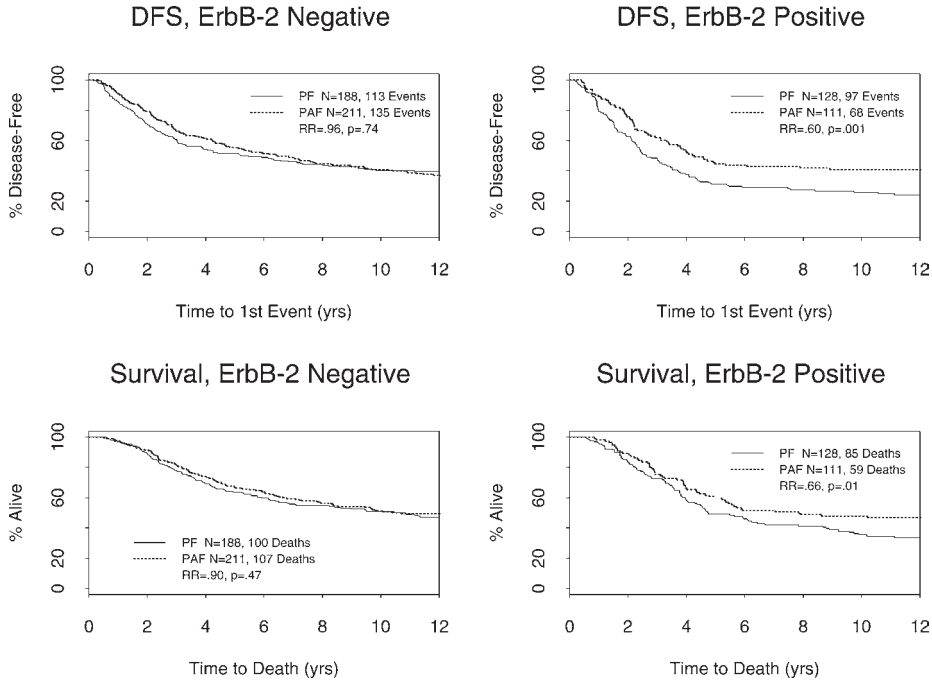

Fig. 1. Kaplan–Meier plots for PF and PAF treatment arms according to erbB-2 status. Two endpoints, disease-free survival (DFS) *(top row)* and survival *(bottom row)*, are shown. Relative risks *(RR)* and *p*-values shown on each plot are from the Cox proportional hazards model with covariates for treatment, age at surgery, clinical tumor size, pathologic lymph node status, and estrogen receptor expression.

cally significant or nearly significant at conventional levels. Similar results were obtained for the RFS and DDFS endpoints.

In the B-11 study, the DFS and survival endpoints were considered primary and the other endpoints secondary, and findings for all four endpoints were consistent. Nevertheless, concern over multiplicity of hypothesis tests prompted the determination of a *p*-value for the interaction effect adjusted for the number of tests. A Bonferroni type adjustment, whereby one multiplies the minimum *p*-value by 4, would constitute an overly conservative adjustment here, because test statistics for the four endpoints are highly correlated. Instead, bootstrap resampling was used to estimate the correlation among the four tests, and the *p*-value associated with the maximum absolute *Z* value was computed by numerical integration. The adjusted *p*-value obtained for the hypothesis of interaction between treatment and erbB-2 status was 0.04.

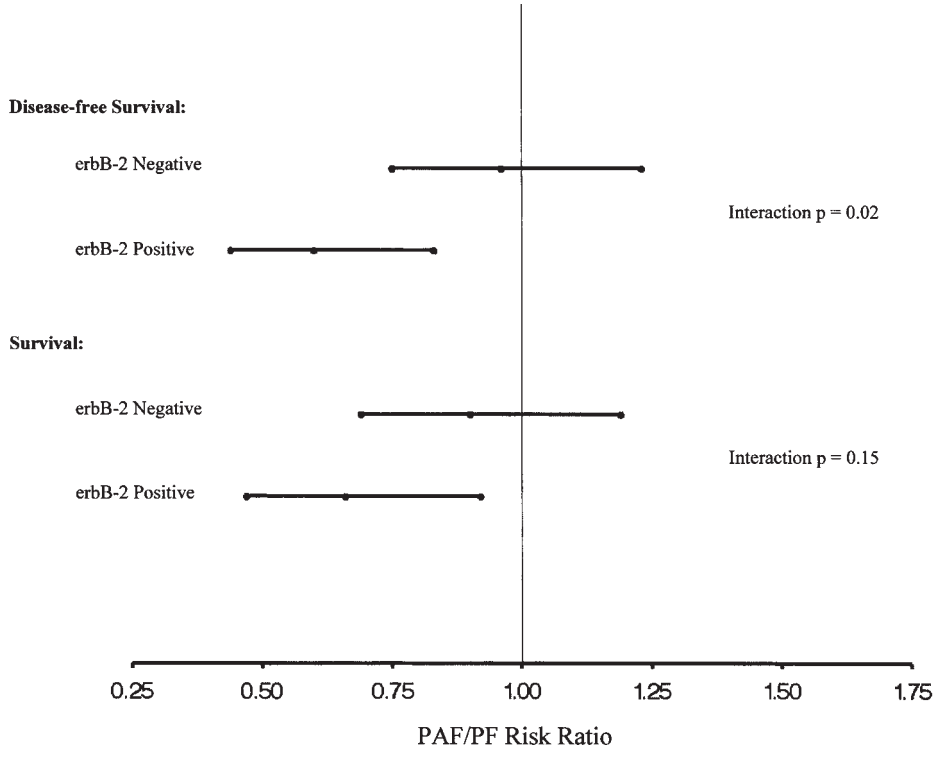

Fig. 2. Relative risk for PAF vs. PF according to erbB-2 status. The *p*-values for interaction between treatment and erbB-2 status are from a Wald test of the cross-product term of these covariates in the Cox proportional hazards model.

4.3. A Prospective Study for erbB-2 *Targeted Therapy*

As a targeted therapeutic strategy, researchers have created a monoclonal antibody to bind to the extracellular domain of the growth factor receptor encoded by the *erbB-2* oncogene, with the goal of inhibiting growth of tumors overexpressing the receptor. Through cell line experiments and animal xenograft models, the antibody was demonstrated to have significant antiproliferative effects. A genetically engineered successor agent (trastuzumab, Herceptin, commercially) was subsequently developed and found to be efficacious in patients with metastatic breast cancer overexpressing erbB-2 ***(49,50)***. The NSABP as well as other clinical trials groups have recently begun trials to evaluate Herceptin in the adjuvant setting. The NSABP trial (protocol B-31) will compare doxorubicin and cyclophosphamide followed by taxol with that same regimen plus Herceptin, in patients with operable erbB-2 positive tumors

and positive lymph nodes. This study will enroll 2700 patients and require roughly 8 yr until definitive results are obtained. Several ancillary investigations concerning erbB-2 are planned, as described in the following paragraphs.

In this multiinstitutional study, participating centers will perform erbB-2 testing using either IHC or FISH and score results as positive or negative according to a common criterion (only patients testing positive will be enrolled). Tumor specimens will also be provided to the central pathology laboratory, where a comprehensive reevaluation of erbB-2 will be conducted in order to address several ancillary study aims. These are: (1) to verify the reported status of the tumor from the institution; (2) to compare results of the various assays, with the opportunity for a direct comparison of methods that measure protein overexpression and those which evaluate gene amplification; and (3) to evaluate whether the assays can predict response to Herceptin. Six assay types will be performed: the DAKO HercepTest kit, TAB250, TAB250/pAb-1 cocktail (used in the B-11 analysis), CB-11, HER-2 FISH assay, and array-based CGH. Several other important pathologic studies will also be performed. One of these involves determining whether expression of the phosphorylated erbB-2 receptor (an indicator that the receptor is capable of binding) in the tumor is prognostic for outcomes or predictive of response to Herceptin, and to evaluate whether this frequency of expression differs in post-relapse tissue among patients either receiving or not receiving Herceptin. Another involves evaluating whether shed extracellular domain of erbB-2 or autoantibodies found in patient serum are associated with outcomes or response to Herceptin.

Explicit hypotheses and analytic methods for these investigations are described in the B-31 protocol document. By the addition of these ancillary studies to this randomized trial designed to evaluate the efficacy of Herceptin in addition to multidrug chemotherapy among erbB-2 positive patients, the B-31 study will address in a comprehensive and prospective manner many of the outstanding research questions concerning erbB-2.

5. Summary and Recommendations for Analyzing Molecular Markers in Clinical Cancer Research

The case study presented here illustrates how previously collected clinical outcomes data, even with a ''retired'' treatment regimen, can serve as a vital resource for advancing the understanding of the natural history of cancer and, furthermore, can play a role in refining treatment selection for current patients. The availability of archived tumor samples allowed for the augmentation of the long-term outcome data under randomized treatment assignment with a modern molecular marker and yielded an important finding of biologic and clinical relevance today. This study, combined with the foundational work that

preceded it and the concurrent results from other large clinical trials groups, have fostered debate regarding breast cancer clinical practice, with erbB-2 evaluation now being advocated by some experts (but not others) as part of a routine clinical evaluation prior to treatment ***(28,51)***. In a recent review and consensus statement regarding prognostic factors in breast cancer from the College of American Pathologists ***(52)***, a detailed list of issues, guidelines, and recommendations related to erbB-2 were provided, many of which were reviewed in this discussion. It is clear that erbB-2 will remain an important research area in breast cancer treatment.

Despite the apparent interaction between erbB-2 and response to doxorubicin-containing chemotherapy regimens, it has yet to be established unequivocally which mechanism is at play in this relationship. In the NSABP study, it can be argued that two factors varied between the PF and PAF groups: the addition of doxorubicin and a simple increase in total chemotherapy dose via the use of three agents instead of two. Similarly, in the CALGB study, the dose of doxorubicin and the other agents varied between treatment groups. Thus, in either study, it can be conjectured that (1) erbB-2 positive tumors may be specifically sensitive to doxorubicin or (2) erbB-2 tumors may be more resistant to chemotherapy and that greater total chemotherapy exposure is beneficial. There is supporting biological information for both mechanisms. Statisticians and other researchers should take heed that, particularly in retrospective studies, observations may be simultaneously consistent with several hypotheses.

To this end, a follow-up investigation by the NSABP has explored further the hypothesis that doxorubicin-containing chemotherapy regimens might specifically be more advantageous in patients with erbB-2 overexpression. Among 2295 eligible node-positive patients entered onto NSABP Protocol B-15, a randomized trial comparing AC, CMF, and a regimen of AC followed by reinduction CMF ***(53)***, 2034 (89%) had immunohistochemical analysis of erbB-2. Statistical analyses were similar to that of B-11, with the primary study hypothesis being whether there was a differential benefit from the doxorubicin-containing regimen (AC) relative to CMF according to erbB-2 status. Findings indicated that the superiority of AC over CMF was restricted to erbB-2 positive patients, although the differences did not reach statistical significance ***(54)***. These results provide further evidence that a regimen containing doxorubicin (or other anthracyclines) is preferred for patients with erbB-2 positive tumors, and unlike the B-11 study, directly addresses current treatment guidelines, as AC and CMF remain in wide use.

In this chapter we have described a retrospective analysis where current clinical data and archived biologic samples were used to address a current question in breast cancer. Despite limitations of the materials and the retro-

spective nature of the investigation, a pragmatic and thoughtful analysis yielded valuable information. With the proliferation of biotechnology, there is an ever-greater need to evaluate markers in cancer. In recent years, statisticians have provided extensive comment on the past and current state of research in marker studies, and have proposed appropriate prospective study designs to improve the quality of research. For example, Simon and Altman describe a study classification scheme similar to that used to describe studies in the evolution of therapeutic agents ***(6)***. Phase I studies are those preliminary studies that establish the potential worth of a given marker. Phase II studies are small, exploratory studies that often demonstrate the usefulness of a marker under less than ideal circumstances. Phase III prognostic factor studies are large, definitive studies that provide a considerable weight of evidence for or against a marker's clinical utility. Their criteria for phase III prognostic factor studies are as follows: First, a valid, reproducible assay will be needed, with documentation of inter- and intra-laboratory variation. Assessors of assay results should be blinded to clinical outcome. The study group should be a well-defined cohort for whom the study referral pattern and eligibility are described. The number of patients subsequently unevaluable for the marker should be small (preferably <15%). Treatment should be standardized for all patients or be randomized. Hypotheses should be stated a priori and include specification of endpoints, definition of scoring for the marker result, identification of patient subsets of interest, and other prognostic factors to be included in the analysis. The number of patients and events should be sufficiently large so that power is adequate for clinically relevant effects. Multiple regression models or stratification methods should be used to establish that the marker is prognostic over and above other known prognostic factors. Confidence limits should be presented for effect measures, with multiplicity of tests taken into consideration.

Statisticians contributing to the summary statement on prognostic factors from the College of American Pathologists ***(4)*** similarly presented a broad and comprehensive view of the design needs of future prognostic factor studies (**Table 2**). These considerations, as well as the comments and recommendations of authors cited throughout this chapter, should serve as a valuable guide for the applied statistician engaged in this important research area.

Acknowledgments

This work was supported by Public Health Service Grants U10-CA-76001 (Dignam), U10-CA-12027 (Paik), and U10-CA-69651 (Dignam, Bryant) from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

Table 2
General Statistical Recommendations from College of American Pathologists Conference XXXV

1. Clinical trials should be specifically designed to test whether a factor has prognostic value. This question can be included in a therapeutic trial, but careful attention must be paid that there is sufficient statistical power to answer both the prognostic and the treatment questions.
2. Prognostic factor question must be prioritized for importance by multidisciplinary groups of investigators working with each cancer type so that the most important factors are quickly evaluated.
3. Journals should adopt publication guidelines for reporting results from prognostic factor studies, including the following elements:
   a. Assessment of possible patient selection bias
      i. Source of patients for the study
      ii. Difference between patients with and without tumor marker in terms of
         1. Baseline demographic and tumor characteristics
         2. Treatment received
         3. Efficacy outcomes
   b. Statement about how missing data were handled
   c. Cut-point selection for method stated
   d. Adjustments for multiple testing
   e. Statistical power analysis if conclusion is negative
   f. Large validation studies should be given publication preference after initial exploratory work for a tumor marker.
4. Organization addressing prognostic factor categorization should come to consensus about the ranking of factors, or at least harmonize their recommendations relative to each other, so that a clear picture of the relative value of various factors is developed.
5. Continued research into multivariate analysis techniques for incomplete data and for evaluation of multiple factors is needed.

Reprinted from **ref. *4*** with permission from the College of American Pathologists.

References

1. McGuire, W. L. and Clark, G. M. (1992) Prognostic factors and treatment decisions in axillary-node-negative breast cancer. *N. Engl. J. Med.* **326,** 1756–1761.
2. Goldhirsch, A., Glick, J. H., Gelber, R. D., and Senn, H. J. (1998) Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. *J. Natl. Cancer Inst.* **90,** 1601–1608.

3. Thomssen, C., Janicke, F., Kaufmann, M., Scharl, A., and Hayes, D. F. (2000) Do we need better prognostic factors in node-negative breast cancer? *Eur. J. Cancer* **36,** 293–306.
4. Hammond, M. E., Fitzgibbons, P. L., Compton, C. C., Grignon, D. J., Page, D. L., Fielding, L. P., et al. (2000) College of American Pathologists Conference XXXV: Solid tumor prognostic factors—which, how, and so what? *Arch. Pathol. Lab. Med.* **124,** 958–965.
5. Hayes, D. F., Bast, R. C., Desch, C. E., Fritsche, H., Jr, Kemeny, N. E., Jessup, J. M., et al. (1996) Tumor marker utility grading system (TMUGS): a framework to evaluate clinical utility of tumor markers. *J. Natl. Cancer Inst.* **88,** 1456–1466.
6. Simon, R. and Altman, D. G. (1994) Statistical aspects of prognostic factor studies in oncology. *Br. J. Cancer* **69,** 979–985.
7. George, S. L. (1994) Statistical considerations and modeling of clinical utility of tumor markers, in *Hematology/Oncology Clinics of North America: Tumor Markers in Adult Solid Malignancies* (Hayes, D. F., ed.) Saunders, Philadelphia, PA, pp. 457–470.
8. Pajak, T. F., Clark, G. M., Sargent, D. J., McShane, L. M., and Hammond, E. H. (2000) Statistical issues in tumor marker studies. *Arch. Pathol. Lab. Med.* **124,** 1011–1015.
9. Hilsenbeck, S. G., Clark, G. M., and McGuire, W. L. (1992) Why do so many prognostic factors fail to pan out? *Breast Cancer Res. Treat.* **22,** 197–206.
10. Altman, D. G., Lausen, B., Sauerbrei, W., and Schumacher, M. (1994) Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J. Natl. Cancer Inst.* **86,** 829–835.
11. Hilsenbeck, S. G. and Clark, G. M. (1996) Practical p-value adjustment for optimally selected cutpoints. *Statist. Med.* **15,** 103–112.
12. Durrleman, S. and Simon, R. (1988) Flexible regression models with cubic splines. *Statist. Med.* **8,** 551–561.
13. Gray, R. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Statist. Assoc.* **87,** 942–951.
14. Bryant, J., Fisher, B., Gunduz, N., Costantino, J. P., and Emir, B. (1998) S-phase fraction combined with other patient and tumor characteristics for the prognosis of node-negative, estrogen-receptor-positive breast cancer. *Breast Cancer Res. Treat*. **51,** 239–253.
15. Cox, D. R. (1972) Regression models and life table. *J. Roy Statist. Soc.* B **34,** 187–220.
16. Schoenfeld, D. A. (1983) Sample-size formula for the proportional-hazards regression model. *Biometrics* **39,** 499–503.
17. Schmoor, C., Sauerbrei, W., and Schumacher, M. (2000) Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statist. Med.* **19,** 441–452.
18. Peterson, B. and George, S. L. (1993) Sample size requirements and length of study for testing interaction in a 2×k factorial design when time-to-failure is the outcome. *Control Clin. Trials* **14,** 511–522. (Erratum [1994] *Control Clin. Trials* **15,** 326).

19. Klinger, A., Donnegger, F., and Ulm, K. (2000) Identifying and modeling prognostic factors with censored data. *Statist. Med.* **19,** 601–615.
20. King, C. R., Kraus, M. H., and Aaronson, S. A. (1985) Amplification of a novel v-*erbB*-related gene in a human mammary carcinoma. *Science* **229,** 974–976.
21. Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., and McGuire, W. L. (1987) Human breast cancer: correlation of relapse and survival with amplification of HER-2/*neu* oncogene. *Science* **235,** 177–182.
22. Slamon, D. J., Godolphin, W., Jones, L. A., Holt, J. A., Wong, S. G., Keith, D. E., et al. (1989) Studies of HER2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244,** 707–712.
23. Press, M. F., Pike, M. C., Chazin, V. R., Hung, G., Udove, J. A., Markowicz, M., et al. (1993) HER-2/neu expression in node-negative breast cancer: direct tissue quantitation by computerized image analysis and association of overexpression with increased risk of recurrent disease. *Cancer Res.* **53,** 4960–4970.
24. Press, M. F., Bernstein, L., Thomas, P. A., Meisner, L. F., Zhou, J. Y., Ma, Y., et al. (1997) HER-2/*neu* gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas. *J. Clin. Oncol.* **15,** 2894–2904.
25. Andrulis, I. L., Bull, S. B., Blackstein, M. E., Sutherland, D., Mak, C., Sidlofsky, S., et al. (1998) *neu/erb*B-2 amplification identifies a poor-prognosis group of women with node-negative breast cancer. *J. Clin. Oncol.* **16,** 1340–1349.
26. Allred, D. C., Clark, G. M., Tandon, A. K., Molina, R., Tormey, D. C., Osborne, C. K., et al. (1992) HER-2/*neu* in node-negative breast cancer: prognostic significance of overexpression influenced by the presence of in situ carcinoma. *J. Clin. Oncol.* **10,** 599–605.
27. Gusterson, B. A., Gelber, R. D., Goldhirsch, A., Price, K. N., Save-Soderborgh, J., Anbazhagan, R., et al. (1992) Prognostic importance of c-erbB-2 expression in breast cancer. International (Ludwig) Breast Cancer Study Group. *J. Clin. Oncol.* **10,** 1049–1056.
28. Clahsen, P. C., van de Velde, C. J., Duval, C., Pallud, C., Mandard, A. M., Delobelle-Deroide, A., et al. (1998) p53 protein accumulation and response to adjuvant chemotherapy in premenopausal women with node-negative early breast cancer. *J. Clin. Oncol.* **16,** 470–479.
29. Piccart, M. J., Di Leo, A., and Hamilton, A. (2000) HER2. a 'predictive factor' ready to use in the daily management of breast cancer patients? *Eur. J. Cancer* **36,** 1755–1761.
30. Press, M. F., Hung, G., Godolphin, W., and Slamon, D. J. (1994) Sensitivity of HER2/neu antibodies in archived tissue samples: potential sources of error immunohistochemical studies of oncogene expression. *Cancer Res.* **54,** 2771–2777.
31. Clark, G. M. (1998) Should selection of adjuvant chemotherapy for patients with breast cancer be based on erbB-2 status? *J. Natl. Cancer Inst.* **90,** 1320–1321.
32. Ravdin, P. M. (1999) Should HER2 status be routinely measured for all breast cancer patients? *Semin. Oncol.* **26,** 117–123.

33. Trock, B. J., Yamauchi, H., Brotzman, M., Stearns, V., and Hayes, D. F. (2000) C-erbB2 as a prognostic factor in breast cancer (BC): a meta-analysis. *Proc. Am. Soc. Clin. Oncol.* **19,** 97.
34. Stal, O., Sullivan, S., Wingren, S., Skoog, L., Rutqvist, L. E., Carstensen, J. M., and Nordenskjold, B. (1995) c-erbB-2 expression and benefit from adjuvant chemotherapy and radiotherapy of breast cancer. *Eur. J. Cancer* **31A,** 2185–2190.
35. Miles, D. W., Harris, W. H., Gillett, C. E., Smith, P., and Barnes, D. M. (1999) 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* **84,** 354–359.
36. Ménard, S., Valagussa, P., Pilotti, S., Biganzoli, E., Boracchi, P., Casalini, P., et al. (1999) Benefit of CMF Treatment in Lymph Node-Positive Breast Cancer Overexpressing HER2. *Proc. Am. Soc. Clin. Oncol.* **17,** 257.
37. Bianco A. R., De Laurentiis, M., Carlomagno, C., Lauria, R., Petrella, G., Panico, L., et al. (1998) 20 year update of Naples GUN trial of adjuvant breast cancer therapy: evidence of interaction between c-erbB-2 expression and tamoxifen efficacy. *Proc. Am. Soc. Clin. Oncol.* **17,** 373.
38. Elledge, R. M., Green, S., Ciocca, D., Pugh, R., Allred, D. C., Clark, G. M., et al. (1998) HER-2 expression and response to tamoxifen in estrogen receptor-negative breast cancer: a Southwest Oncology Group study. *Clin. Cancer Res.* **4,** 7–12.
39. Berry, D. A., Muss, H. B., Thor, A. D., Dressler, L., Liu, E. T., Broadwater, G., et al. (2000) HER-2/neu and p53 expression versus tamoxifen resistance in estrogen receptor–positive, node-positive breast cancer. *J. Clin. Oncol.* **18,** 3471–3479.
40. Muss, H. B., Thor, A. D., Berry, D. A., Kute, T., Liu, E. T., Koerner, F., et al. (1994) c-erbB-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *N. Engl. J. Med.* **330,** 1260–1266. (Erratum [1994] *N. Engl. J. Med.* **331,** 211).
41. Thor, A. D., Berry, D. A., Budman, D. R., Muss, H. B., Kute, T., Henderson, I. C., et al. (1998) erbB-2, p53, and the efficacy of adjuvant therapy in lymph node-positive, hormone receptor-negative breast cancer. *J. Natl. Cancer Inst.* **90,** 1346–1360.
42. Ravdin, P. M., Green, S., Albain, K. S., Boucher, V., Ingle, J., Pritchard, K., et al. (1998) Initial report of the SWOG biologic correlation study of c-erbB-2 expression as a predictor of outcome in a trial comparing adjuvant CAFT to tamoxifen alone. *Proc. Am. Soc. Clin. Oncol.* **17,** 374.
43. Paik, S., Bryant, J., Park, C., Fisher, B., Tan-Chiu, E., Hyams, D., et al**.** (1998) erbB-2 and response to doxorubicin in patients with axillary lymph node-positive, hormone receptor-negative breast cancer. *J. Natl. Cancer Inst.* **90,** 1361–1370.
44. Pietras, R. J., Fendly, B. M., Chazin, V. R., Pegram, M. D., Howell, S. B., and Slamon, D. J. (1994) Antibody to HER-2/neu receptor blocks DNA repair after cisplatin in human breast and ovarian cancer cells. *Oncogene* **9,** 1829–1838.
45. Paik, S., Hazan, R., Fisher, E. R., Sass, R. E., Fisher, B., Redmond, C., et al. (1990) Pathologic findings from the National Surgical Adjuvant Breast and Bowel Project: prognostic significance of erbB-2 protein overexpression in primary breast cancer. *J. Clin. Oncol.* **81,** 103–112.

46. Fisher, B., Redmond, C., Wickerham, D. L., Bowman, D., Schipper, H., Wolmark, N., et al. (1989) Doxorubicin containing regimens for the treatment of stage II breast cancer: the National Surgical Breast and Bowel Project experience. *J. Clin. Oncol.* **7,** 572–582.
47. Nelson, N. J. (2000) Experts debate value of HER2 testing methods [News]. *J. Natl. Cancer Inst.* **92,** 292–294.
48. Mitchell, M. S. and Press, M. F. (1999) The role of immunohistochemistry and fluorescence in situ hybridization for HER2/neu in assessing the prognosis of breast cancer. *Semin. Oncol.* **26,** 108–116.
49. Cobleigh, M. A., Vogel, C. L., Tripathy, D., Robert, N. J., Scholl, S., Fehrenbacher, L., et al. (1999) Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. *J. Clin. Oncol.* **17,** 2639–2648.
50. Norton, L., Slamon, D., Leyland-Jones, B., Wolter, J., Fleming, T., Eirmann, W., et al. (1999) Overall survival (os) advantage to simultaneous chemotherapy (CRx) plus the humanized anti-HER2 monoclonal antibody Herceptin (H) in HER2-overexpressing (HER2+) metastatic breast cancer (MBC). *Proc. Am. Soc. Clin. Oncol.* **18,** 483.
51. Hayes, D. F., Yamauchi, H., Stearns, V., Brotzman, M., Isaacs, C., and Trock B. (2000) Should all breast cancers be tested for c-erbB-2? in *2000 ASCO Educational Book* (Perry, M. C., ed.). American Society of Clinical Oncology, Alexandria, VA, pp. 257–265.
52. Fitzgibbons, P. L., Page, D. L., Weaver, D., Thor, A. D., Allred, D. C., Clark, G. M., et al. (2000) Prognostic factors in breast cancer: College of American Pathologists Consensus Statement 1999. *Arch. Pathol. Lab. Med.* **124,** 966–978.
53. Fisher, B., Brown, A. M., Dimotrov, N. V., Poisson, R., Redmond, C., Margolese, R. G., et al. (1990) Two months of doxorubicin-cyclophosphamide with and without interval reinduction therapy compared with 6 months of cyclophosphamide, methotrexate, and fluorouracil in positive-node breast cancer patients with tamoxifen-nonresponsive tumors: results from the National Surgical Adjuvant Breast and Bowel Project B-15. *J. Clin. Oncol.* **8,** 1483–1496.
54. Paik, S., Bryant, J., Tan-Chiu, E., Yothers, G., Park, C., Wickerham, D. L., and Wolmark, N. (2000) HER2 and choice of adjuvant chemotherapy for invasive breast cancer: NSABP Protocol B-15. *J. Natl. Cancer Inst.* **92,** 1991–1998.