The Use of Genomics in Clinical Trial Design

Richard Simon

Abstract Many cancer treatments benefit only a minority of patients who receive them. This results in an enormous burden on patients and on the health care system. The problem will become even greater with the increasing use of molecularly targeted agents whose benefits are likely to be more selective unless the drug development process is modified to include codevelopment of companion diagnostics. Whole genome biotechnology and decreasing costs of genome sequencing make it increasingly possible to achieve an era of predictive medicine in oncology therapeutics. The challenges are numerous and substantial but are not primarily technological. They involve organizing publicly funded diagnostics of deregulated pathways, adopting new paradigms for drug development, and developing incentives for industry to incur the complexity and expense of codevelopment of drugs and companion diagnostics. This article reviews some designs for phase III clinical trials that may facilitate movement to a more predictive oncology.

Most cancer treatments benefit only a minority of patients to whom they are administered. This results in adverse events and substantial health care costs for treatment of cancer patients who receive no benefit. Accumulating understanding of genomic differences among tumors of the same primary site indicates that most molecularly targeted agents are likely to benefit only the patients whose tumors are driven by deregulation of the targeted pathways. It is important that new drugs be developed with companion diagnostics that identify the patients who are most likely to benefit from the new regimen. It is often very difficult to determine this after the treatment is in broad use. Successful prospective codevelopment of a drug and companion diagnostic presents many new challenges, however. In this article, we will address some of the issues in the design of phase III clinical trials for new treatments and diagnostic tests that may indicate which patients benefit from the new treatment. We will not discuss in detail the early steps in the development of the diagnostic.

Biomarkers

The term "biomarker" can be used for a wide variety of purposes and this often leads to confusion in discussion of biomarker development, use, and validation. Traditionally, a biomarker was a measurement that tracks the pace of a disease, increasing as the disease progresses and decreasing as it regresses. There are many potential uses for such biomarkers for measuring antitumor effect in phase I and phase II studies conducted to establish proof of concept and identify an appropriate dose. For such uses, the biomarker need not be a validated surrogate of clinical benefit, only a measure of treatment effect in which the developers have sufficient confidence to use in preparing the way for a phase III trial. Claims of treatment benefit should rarely be made based on phase I or II trials. The standards for validation of a biomarker as a surrogate of clinical benefit are stringent (1, 2). It is not sufficient to show that the biomarker is correlated with clinical outcome. Partial tumor regression is generally not a valid surrogate for survival although responders often survive longer than nonresponders. Validation of a biomarker as a surrogate of clinical benefit generally requires a series of randomized clinical trials that show that treatment versus control arm differences with regard to average biomarker change are concordant with treatment versus control arm differences with regard to clinical outcome (2, 3).

Biomarkers can also play key roles in prognostic and predictive characterizations of a patient's disease. Prognostic markers are baseline measurements that provide information about the patients' likely long-term outcome either untreated or with standard treatment. Prognostic markers can be used to determine whether the patient requires any systematic treatment or any cytotoxic chemotherapy. Predictive markers are baseline measurements that indicate whether the patient is a good candidate for a specific drug or regimen. Biomarker diagnostics used for these purposes can be powerful tools for improving patient management and for enhancing the effectiveness of clinical development. The purpose of prognostic and predictive biomarkers is completely different from the purpose of biomarkers used as surrogate end points. Because "validation" or "qualification" only has meaning in terms of fitness for the intended use, the criteria for validation of surrogate end points should not be mistakenly applied to prognostic or predictive biomarkers (4). The validation of prognostic and predictive biomarkers, although demanding, is often much more feasible than the validation of biomarkers as surrogate

Authors' Affiliation: National Cancer Institute, Bethesda, Maryland
Received 2/15/08; revised 6/18/08; accepted 6/20/08.
Requests for reprints: Richard Simon, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434. Phone: 301-496-0975; Fax: 301-402-0560; E-mail: rsimon@nih.gov.

doi:10.1158/1078-0432.CCR-07-4531

end points. The Food and Drug Administration guidance for industry on pharmacogenomic data submissions seems to be written based on experience with biomarkers intended for use as surrogate end points but proposed as applying to all types of biomarkers (5).

Prognostic and Predictive Classifiers

The oncology literature is replete with publications on prognostic factors but very few of these are used in clinical practice. For example, Pusztai et al. (6) identified 939 publications over a 20-year period on prognostic factors in breast cancer, but only four factors [estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2 (HER2), urokinase plasminogen activator, and Oncotype DX] were recommended for use by the American Society of Clinical Oncology (7). Prognostic factors are rarely used unless they help with therapeutic decision making. Most prognostic factor studies are conducted using a convenience sample of patients whose tissues are available (8). Often these patients are too heterogeneous with regard to treatment, stage, and standard prognostic factors to support therapeutically relevant conclusions (9). Many publications attempt to show that new factors are "independently prognostic" or are more prognostic than standard factors, but these analyses often fail to identify a role of the new factors in therapeutic decision making. Hayes et al. (10, 11) have developed and used a tumor marker utility grading system to assist in the evaluation of the clinical utility of tumor markers. The REMARK guidelines will hopefully promote better study design for the development of prognostic markers (12).

Predictive biomarkers identify patients who are likely or unlikely to benefit from a specific treatment. For example, *HER2* amplification is a predictive classifier for benefit from trastuzumab and perhaps also from doxorubicin (13, 14) and paclitaxel (15). The presence of a mutation in epidermal growth factor receptor (EGFR) may be a predictive marker for response to EGFR inhibitors (16), although it is unclear today whether *EGFR* amplification is a better predictive marker or whether either is sufficiently predictive for clinical use (17). A predictive biomarker may be used to identify patients who are poor candidates for a particular drug; for example, colorectal cancer patients whose tumors have KRAS mutations may be poor candidates for treatment with EGFR inhibitors (18).

Prognostic or predictive biomarkers are often based on the sequence, copy number, translocation, phosphorylation, methylation, or expression of a single gene. Expression may be measured either at the mRNA transcript or protein level. Such single gene/protein biomarkers are attractive because they are often closely linked to the mechanism of action of the drug and are thus biologically interpretable. In some cases, the target of the drug is known but it is not clear how to best measure the essentiality of the target to the pathogenesis of a specific tumor. For example, although trastuzumab was initially developed using a test for protein expression of HER2, subsequent classification has often been based on the amplification of the gene (19). In other cases, the target of the drug may not be clearly known and the options for measurement will be more numerous. If a diagnostic is to be codeveloped with the drug, the phase II studies must be designed to evaluate the candidate assays available, to select one, and then to perform analytic validation of the assay before launching the phase III trial. Analytic validation refers to establishment of robustness and reproducibility of the assay and measurement of sensitivity and specificity relative to a "gold standard" assay if one is available (4).

We will use the term "classifier" to refer to a diagnostic that translates one or more biological measurements to a set of predicted categories. For example, with a prognostic classifier, the categories may refer to low risk of tumor recurrence, moderate risk of recurrence, and high risk of recurrence. With a predictive classifier, the categories may refer to patients most likely to benefit from the new regimen and less likely to benefit. A biomarker based on a measurement involving a single gene or protein can be converted to a classifier by introducing one or more cutoff points, depending on how many categories are desired. Classifiers can also be defined based on summary measures of combinations of many variables, as in the case of gene expression profiling (20).

Prognostic classifiers based on gene expression data should be developed in a manner that addresses a focused therapeutic decision context. For example, the Oncotype DX recurrence index was developed by studying women with breast cancer whose tumors were estrogen receptor positive, had not spread to the axillary lymph nodes, and who had received tamoxifen as the only systemic therapy (21, 22). A score was developed based on tumor expression of 21 genes to identify women whose disease-free survival was sufficiently good that they might elect to forgo cytotoxic therapy. Prognostic factors developed in such a focused manner can be relevant for therapeutic decisions. The score is often used as a classifier by introducing two cutoff points to distinguish patients with low, moderate, and high risks of tumor recurrence.

For developing a predictive classifier of patients likely to benefit from a new drug, one can perform gene expression profiling of patients on phase II trials of the drug and compare the expression profiles of responders to those of nonresponders. In that way, one can identify the differentially expressed genes and determine how to combine and weight expression levels for the component genes and to establish a cutoff point that optimizes predictive accuracy of the classifier. Dobbin et al. (23, 24) have developed methods for planning the number of cases needed to effectively develop such a classifier. Larger phase II studies may be required to have sufficient responders in the phase II database for this approach (25).

There is substantial literature on the development of gene expression–based classifiers (e.g., refs. 26–32) and the process is too extensive to be reviewed here. It is important to emphasize that important components of the process include identification of the genes to be included in the classifier, selecting a mathematical way of combining the expression levels of the individual genes, and training the classifier (i.e., determining the weights and cutoff points) on a training set of data to distinguish responders from nonresponders (33). A gene expression classifier is not just a set of genes. The process of building a classifier based on genome-wide expression profiling is very different than traditional statistical model building. Because the number of variables (genes) available is

much greater than the number of cases available in the training data set, traditional statistical regression model building strategies are ineffective. Traditional approaches that ensure that all variables included in the model contribute to prediction and that their interaction with other variables is properly modeled often result in over-fitted models that predict poorly. With high-dimensional data, one must focus clearly on the objective of accurate prediction and not confuse this objective with that of achieving biological insight or ensuring that all variables included are essential, or that the model is "correct" (34). Criticisms of models that predict accurately based on grounds that they do not provide biological insight or that not all of the components may be essential are misguided. As with physics, predictive models can have great power and benefit, even if they do not clarify the nature of reality. It is possible to develop models that predict accurately when the number of cases is fewer than the number of variables, if recent methods of model development and validation are used.

The BRB-ArrayTools software provides extensive resources for development of a wide range of prognostic and predictive classifiers based on gene expression data for binary response or survival end points (35). It provides an environment for developing a classifier on a training set and estimating the accuracy of the model on a test set of data or for using a wide range of resampling methods for estimating the predictive accuracy of the model developed on the complete set of data. All of these methods represent internal clinical validation (34, 36). Comparison of the split sample approach and resampling methods has been done by Molinaro et al. (37) in the context of studies where the number of variables is much greater than the number of cases. There are generally substantial biases entailed in using the same data for model development and testing without proper use of resampling (34, 38), and these biased reports are prevalent in the oncology literature (39). BRB-ArrayTools is available for downloading online.[1] Important issues in the preprocessing of microarray data are discussed by Owzar et al. (40) in this series.

In this article, we will focus on the use of predictive classifiers in the design of phase III trials to determine whether a new drug is effective and how its effectiveness relates to the classes defined by a predictive classifier. In some cases, we will use the phrase diagnostic test to cover predictive classifiers defined either by single genes/proteins or predictive classifiers based on combining the expression levels of multiple genes in defined manners with defined cutoff points to establish predictive classes.

Enrichment Design

The objective of a phase III pivotal clinical trial is to evaluate whether a new drug, given in a defined manner, has medical utility for a defined set of patients. Pivotal trials test prespecified hypotheses about treatment effectiveness in specified patient population groups. The role of a predictive biomarker classifier is to specify the population of patients. The process of classifier development may be exploratory and subjective, but the use of the classifier in the pivotal trial must not be.

Figure 1 shows a design in which a diagnostic test is used to define eligibility for a randomized clinical trial comparing a new drug to a control regimen. This approach was used for the development of trastuzumab. Patients with metastatic breast cancer whose tumors expressed HER2 at a 2+ or 3+ level based on an immunohistochemistry test were eligible for randomization (41). The clinical trial randomized 469 patients but the number of patients whose tumors were tested was not stated. If 75% of patients had available specimens and adequate tests and 25% of patients with adequate tests were HER2 positive, then ~2,500 patients would have to be screened to obtain 469 eligible for randomization.

Simon and Maitournam (42–44) studied the efficiency of this approach relative to the standard approach of randomizing all patients without measuring the diagnostic. They found that the efficiency of the enrichment design depended on the prevalence of test-positive patients and on the effectiveness of the new treatment in test-negative patients. For binary end point trials, they showed that the ratio of number of patients to be randomized for the standard trial (n_S) compared with the number randomized in the enrichment trial (n_E) is approximately

$$n_S/n_E \approx f/(\text{prev} + (1 - \text{prev})\delta_-/\delta_+)^2 \qquad \text{(A)}$$

where prev is the proportion of patients who are test positive; δ_- is the treatment effectiveness for test-negative patients; and δ_+ is the treatment effect for test-positive patients. The variable f is a constant that does not depend on the prevalence or treatment effects; it is generally close in value to 1 unless the control response rate is very low. In cases where the new treatment is completely ineffective in test-negative patients, the formula above simplifies to approximately $1/\text{prev}^2$. Often, however, it is unrealistic to expect that the treatment will be completely ineffective for test-negative patients. The treatment may have some effectiveness for test-negative patients either because the assay is imperfect for measuring deregulation of the putative molecular target or because the drug has off-target effects. Because these alternatives cannot generally be distinguished, there is little value in decomposing δ_- into these components. However, because of the limited specificity of the test, δ_- may not be zero. If the new treatment is half as effective in test-negative patients as in test-positive patients, then the right-hand side of Eq. A simplifies to approximately $4/(\text{prev} + 1)^2$. This equals ~2.56 when 25% of the patients are test positive, indicating that the enrichment design reduces the number of required patients to randomize by a factor of 2.56.

To obtain n_E test-positive patients for randomization, one must screen approximately n_E/prev patients. Simon and Maitournam also compared the enrichment design to the standard design with regard to the number of screened patients. These methods of sample size planning for the design of enrichment trials are available online.[2] The web-based programs are available for binary, survival or disease-free survival,

[1] http://linus.nci.nih.gov/brb

[2] http://linus.nci.nih.gov/brb

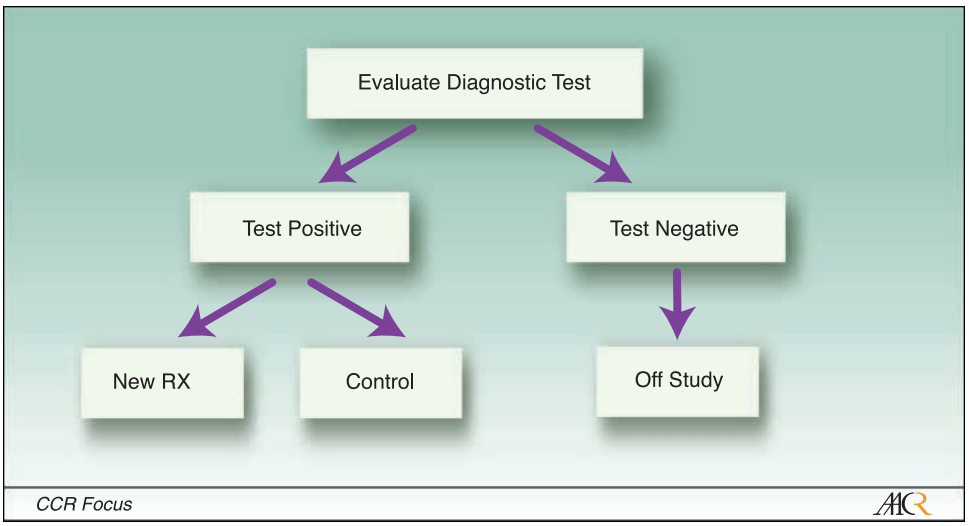

Fig. 1. Flow diagram of the enrichment design in which a prognostic or predictive classifier is used to restrict eligibility to a randomized phase III clinical trial comparing a new treatment to a control regimen. Tissue is obtained from all consenting patients who would be eligible for the corresponding phase III trial if the classifier was not used. The binary classifier is measured from the collected specimen. The binary classifier may be based on a single gene or protein or may be based on measurements of large number of genes or proteins as discussed in "Prognostic and Predictive Classifiers." In either case, the threshold for test positivity is determined in advance. Patients who are classifier positive are then randomized to receive the new treatment or the control regimen. Classifier-negative patients are not treated on the study.

or uncensored quantitative end points. The planning takes into account the performance characteristics of the tests and specificity of the treatment effects. The programs are easy to use and provide comparisons to standard nonenrichment designs based on the number of randomized patients required and the number of patients needed for screening to obtain the required number of randomized patients.

When fewer than half of the patients are test positive and the new treatment is relatively ineffective in test-negative patients, the number of randomized patients required for an enrichment design is often dramatically smaller than the number of randomized patients required for a standard design. This was the case for trastuzumab. The enrichment design that led to approval of trastuzumab was conducted in 469 patients with metastatic breast cancer whose tumors overexpressed HER2 based on immunohistochemical analysis in a central laboratory. The results were highly significant with regard to several end points including 1-year survival rate (78% versus 67%). The trial of 469 patients provides 90% power for detecting a 13.5% improvement in the 1-year survival rate above a baseline of 67% with a two-sided 5% significance level. If benefit from the drug was limited to the 25% of patients expected to be test positive, then the overall improvement in 1-year survival rate would be only 3.375% for a standard design and a total of ~8,050 patients would be required for 90% power to detect such a small effect. This is 17.2 times as many patients as for the enrichment design, in good agreement with the ratio of 16 computed from the approximate form of Eq. A with $f = 1$. If the test-negative patients benefit half as much as the test-positive patients, then 1,254 total patients would be required for 90% power with the standard design. This is 2.7 times as many for the enrichment design but is much less than the number required for screening with the enrichment design.

Focusing initial development on test-positive patients can lead to clarity in determining who benefits from the drug. If the enrichment design establishes that the drug is effective in test-positive patients, the drug could be later developed in test-negative patients. This is preferable to testing new drugs in heterogeneous populations resulting in false-negative results for the overall population.

The enrichment design is particularly appropriate for contexts where there is such a strong biological basis for believing that test-negative patients will not benefit from the new drug that including them would be unethical. In many situations, the biological basis is not compelling. Although our understanding of the molecular target of a drug is often flawed, we do not really want to be including test-negative patients in a clinical trial to show that a treatment that we do not believe will work for them actually does not work. This is particularly true when the drug has adverse effects. If the treatment is shown to be effective in test-positive patients and if there is a robust assay for the test, then it could be argued that medical utility has been shown for the new treatment and for the test. If, however, the test requires approval as a medical device, then either biological or empirical evidence that the new drug is not effective in test-negative patients will be required. In some cases, this might be achieved using data from phase II single arm studies that did not restrict entry based on the classifier.

Including Both Test-Positive and Test-Negative Patients

Instead of using the predictive classifier as an exclusion criterion, both test-positive and test-negative patients may be included and randomized between the new treatment and control groups as indicated in Fig. 2 (45–47). It is essential that an analysis plan be predefined in the protocol for how the predictive classifier will be used in the analysis. It is not sufficient to just stratify the randomization with regard to the classifier without specifying a complete analysis plan. In fact, for many statisticians, stratification of the randomization is not essential for inference; its main importance is that it ensures that only patients with adequate specimens and interpretable test results will enter the trial.

It is important to recognize that the purpose of this design is to evaluate the new treatment in the subsets determined by the

prespecified classifier. The purpose is not to modify the classifier. If the classifier is a composite gene expression–based classifier, the purpose of the design is not to reexamine the contributions of each component. If one makes any such changes, then an additional phase III trial may be needed to evaluate treatment benefit in subsets determined by the new classifier. In moving from post hoc correlative science to reliable predictive medicine, it is important to strictly separate the data used for developing classifiers from the data used for testing treatment effects in subsets determined by those classifiers. Only by honoring this principle can reliable conclusions be achieved. The process of classifier development can be exploratory, but the process of evaluating treatments should not be; it should be based on testing prespecified hypotheses in prespecified patient groups.

Analysis plan: analysis of test negatives contingent on significance in test positives. The simplest analysis plan would consist of separate comparisons of the new treatment to the control in the test-positive and test-negative patients. In cases where a priori one does not expect the treatment to be effective in the test-negative patients unless it is effective in the test-positive patients, one might structure the analysis in the following manner: Compare treatment versus control in test-positive patients using a threshold of significance of 5%. If the treatment difference in test-positive patients is not significant, do not perform statistical significance test in negative patients. Otherwise, compare treatment to control in the test-negative patients using a threshold of statistical significance of 5%. This sequential approach controls the overall type I error at 5%.

With this analysis plan, the number of test-positive patients required is the same as for the enrichment design, denoted n_E above. When that number of patients are accrued, there will be approximately n_E/prev total patients and approximately n_- = (1 - prev) n_E/prev test-negative patients. One should make sure that the n_E is large enough that there is expected to be an adequate number of test-negative patients for analysis. With a time-to-event end point like survival or disease-free survival, the planning will be somewhat more complex. For example, suppose we wish to have 90% power in the test-positive patients for detecting a 50% reduction in hazard for the new treatment versus control at a two-sided 5% significance level. This requires ~88 events of test-positive patients. At the time that there are E_+ events in test-positive patients, there will be approximately E_- events in the test-negative patients and they are related by:

$$E_- = E_+ \left(\frac{\gamma_-}{\gamma_+}\right)\left(\frac{1-\text{prev}}{\text{prev}}\right). \qquad \text{(B)}$$

In Eq. B the symbols λ_- and λ_+ denote the event rates in the test-negative and test-positive control groups at the time that there are E_+ events in the test-positive group. If the control group event rates at that time are the same in the test-negative and test-positive strata, then the ratio of λ's in Eq. B will a have value of 1. If E_+ is 88, if the prevalence of test-positive patients is 25%, and if the control group event rates are equal, then E_- will be ~264 at the time of analysis. This will provide ~90% power for detecting a 33% reduction in hazard at a two-sided significance level of 5%. This power calculation is conditional on the expected number of events, however. To control the power more adequately, the time of analysis of the test-negative patients should be delayed until this number of events is obtained in that subset. On average, the trial will not be delayed compared with the enrichment design, but a large number of test-negative patients will be randomized, treated, and followed up on the study rather than excluded as for the enrichment

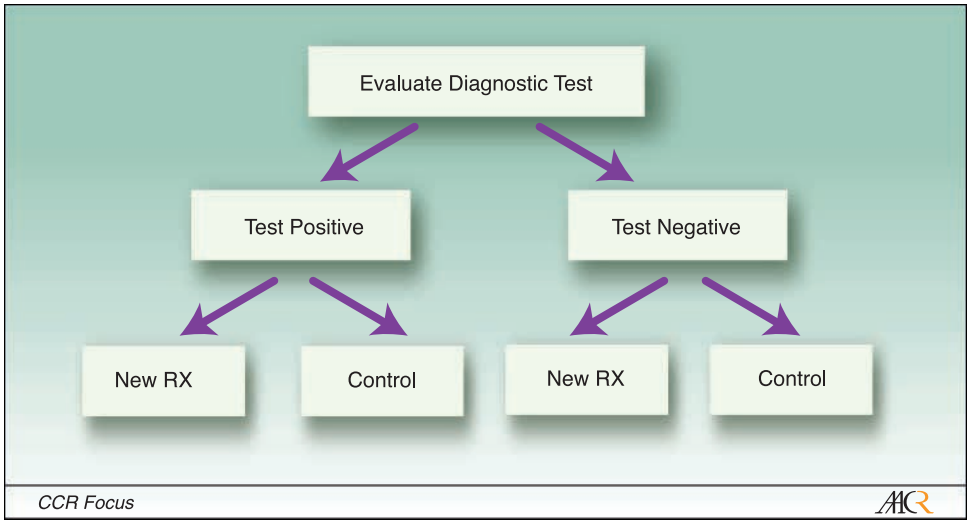

Fig. 2. Flow diagram in which a predictive classifier is used to structure the primary analysis of a randomized phase III clinical trial comparing a new treatment to a control regimen. Tissue is obtained from all consenting patients who would be eligible for the corresponding phase III trial if the classifier was not used. The binary classifier is measured from the collected specimen. The binary classifier may be based on a single gene or protein or may be based on measurements of large number of genes or proteins as discussed in "Prognostic and Predictive Classifiers." In either case, the threshold for test positivity is determined in advance. Both classifier-positive and classifier-negative patients are randomized to receive either the new treatment or control regimen. Stratifying the randomization to ensure balance is recommended, but what is most essential is that tissue and an adequate test result are obtained from all patients and that a primary statistical analysis plan is established that describes exactly how the classifier results will be used in the treatment comparison. Several such analysis plans are described in "Including Both Test-Positive and Test-Negative Patients".

design. If the proportion of test-positive patients was 50% instead of 25%, then at the time that there are 88 events in test-positive patients, there will be about an equal number of events in test-negative patients, assuming that control group event rates are equal. Hence, one will not have 90% power for detecting a 33% reduction in the hazard in test-negative patient; one will only have 90% power for detecting a 50% reduction. It is possible that at that time, a futility analysis in the test-negative patients will indicate that there is unlikely to be a medically important treatment effect in test-negative patients. If the futility analysis does not support such a conclusion, one should continue to follow up test-negative patients and perform the final analysis at a time when there are substantially more events. This may necessitate planning the trial to accrue continually until there are 88 events in the test-positive group and possibly continuing accrual of test-negative patients if the results are significant in the positive group.

Gefitinib is a small-molecule inhibitor of epidermoid growth factor receptor (EGFR) kinase activity. Two untargeted phase III trials of standard chemotherapy with or without gefitinib in chemotherapy naïve patients with advanced non–small-cell lung cancer failed to show a benefit of gefitinib (48, 49). Provocative reports based on specimens from phase II trials indicated that response to gefitinib alone in previously treated patients can be predicted on the basis of somatic mutations in the tyrosine kinase domain of the *EGFR* gene but that the mutations are prevalent in only about 10% to 15% of western patients (50, 51). A randomized placebo controlled clinical trial of erlotinib, another EGFR inhibitor, in unselected patients with non–small-cell lung cancer who had progression after standard chemotherapy showed a statistically significant effect on overall survival (52). Post hoc analysis of available specimens from patients on that clinical trial suggested that the benefit may have been limited to those with high polysomy or amplification of *EGFR* but specimens were available for a minority of patients (53). To clarify this situation, a multicenter randomized clinical trial of gefitinib in non–small-cell lung cancer is being planned that will stratify but not preselect patients based on EGFR. The planning of this trial is complicated by the presence of three tests (protein expression of EGFR, mutation, and amplification of the gene) and the differences in prevalence of positivity among them. Using the methods described here, one might size this trial in the following manner. To be able to detect a 33% reduction in hazard for the use of gefitinib for fluorescence *in situ* hybridization (FISH)–positive cases with 90% power at a 5% two-sided significance level would require ~263 events. Although amplification of *EGFR* seems to be a good prognostic feature, the prognosis of patients with advanced non–small-cell lung cancer is sufficiently poor that the ratio of event rates at final analysis in Eq. B can be assumed to be ~1. If the FISH test is positive in ~30% of cases, then Eq. B indicates that when there are 263 events in FISH-positive cases, there will be, on average, 614 events in FISH-negative cases. This would provide ~90% power for detecting a 23% reduction in hazard for the use of gefitinib in FISH-negative patients at a two-sided 5% significance level. Accrual of 1,200 eligible consented patients for testing would provide ~960 available for randomization if the assay success rate is 80%. Of these 960 patients, ~30%, or 288, would be expected to be positive for amplification of *EGFR* as assessed by FISH, and ~70%, or 672, negative. At the time that the event rate is 90% (the progression free survival of patients with previously treated metastatic disease is short), there will be ~259 available events in the FISH-positive stratum and ~604 events in the FISH-negative stratum, close to the planned event targets.

Analysis plan: analysis determined by interaction test. The traditional statistical approach to the analysis of data in which cases are classified by treatment and by a binary covariate that may effect treatment efficacy is to first test whether there is a significant interaction between treatment efficacy (treatment versus control) and the covariate (test negative and positive). The interaction test is often done at a threshold (α_i) above the traditional 5% level. If the interaction test is not significant, then the treatment effect is evaluated overall, not within levels of the covariate. If the interaction test is significant, then treatment effect is evaluated separately within the levels of the covariate (e.g., test-positive and test-negative classes). This is similar to the test proposed by Sargent et al. (46). In the example described above with 88 events in test-positive patients and 264 events in test-negative patients, the interaction test will have ~93.7% power at a one-sided significance level of 0.10 for detecting an interaction with 50% reduction in hazard for test-positive patients and no treatment effect in test-negative patients. If the treatment also reduces the hazard by 33% in test-negative patients, the interaction test has little power, but that is fine because the overall analysis of treatment effect will be appropriate in that circumstance.

Computer simulations indicate that the two-stage analysis plan with $\alpha_i = 0.10$ has the following properties. With 88 test-positive patients and 264 test-negative patients, the design detects a significant interaction and detects a significant treatment effect in test-positive patients in 88% of replications when the treatment reduces hazard by 50% in test-positive patients and is ineffective in test-negative patients. If the treatment reduces hazard by 33% in both test-positive and test-negative patients, the interaction is nonsignificant and the overall treatment effect is significant in 87% of cases. The overall treatment effect refers to the comparison of treatment to control that includes both test-negative and test-positive patients.

If one were planning a trial to detect a uniform 33% reduction in hazard with 90% power and 5% two-sided significance level, one would require ~256 events. If 25% of the cases were test positive and the control group event rates in test-negative and test-positive patients are about the same, then at time of analysis there would be ~64 events in test-positive cases and 192 events in test-negative cases. If the treatment reduces hazard by 33% in both test-positive and test-negative patients, the interaction is nonsignificant and the overall treatment effect is significant in ~81% of cases. If the treatment reduces hazard by 50% in test-positive cases and is ineffective in test-negative cases, then the interaction is significant and the treatment effect in test-positive cases is significant in 76% of replications. Thus, even if the trial is sized for detecting a uniform 33% reduction in hazard, the two-stage analysis plan will have ~76% power for detecting a substantial treatment effect restricted to the test-positive patients.

Other analysis plans. Simon and Wang (47) proposed an analysis plan in which the new treatment group is first compared with the control group overall. If that difference is not significant at a reduced significance level α_1 (e.g., 0.03), then the new treatment is compared with the control group just for test-positive patients. The latter comparison uses a threshold of significance of $\alpha - \alpha_1$ (e.g., 0.02), or whatever portion of the total α (e.g., 0.05) is not used by the initial test. This design emphasizes the overall comparison and provides an incentive for the sponsor to develop a predictive classifier as a backup strategy for drug approval. The sample size needs to be set to provide adequate power for the overall test at the reduced significance level α_1 and for the potential subset analysis at level $\alpha - \alpha_1$. The targeted treatment effect within the subset would, however, be greater than the overall treatment effect used for power calculations. This design is also discussed by Song and Chi (54) and by George (34) using a refinement of the significance levels that takes into account the correlation between the test of overall treatment effect and the treatment effect within the test-positive subset.

An analysis plan can also be constructed based on the use of a test statistic to indicate whether the strongest evidence of treatment effectiveness occurs in the test-positive patients or for the overall group. Let the estimated treatment effects for test-positive and test-negative patients be denoted $\hat{\delta}_+$ and $\hat{\delta}_-$, respectively. They have variances approximately $4/E_+$ and $4/E_-$ respectively. An overall estimate of treatment effect $\hat{\delta}$ having variance 1 can be constructed as a weighted average of these subset-specific estimates with the weights equal to the proportion of events in the subsets. Then a test statistic

$$z_{\min} = \min\{\hat{\delta}, \frac{\hat{\delta}_+}{\sqrt{4/E_+}}\}$$

indicates the strongest evidence of treatment effectiveness. The minimum is used in the above formula because a negative log hazard ratio indicates effectiveness of the new treatment compared with control. The critical value for assessing the statistical significance of $z_{\min}$ can be determined based on the approximate bivariate null distribution of the two components or based on permutation of the treatment group labels of the cases.

With 88 test-positive events and 264 test-negative events, the statistical power of the $z_{\min}$ test is 86.7% if the new treatment reduces the hazard by 50% in test-positive cases and is ineffective for test negatives. The statistical power is 95.1% if the new treatment reduces the hazard by 33% for both test-positive and test-negative cases. If the trial were planned for a total of only 256 events, 64 test positive and 192 test negative, then the statistical power is 70% if the new treatment reduces the hazard by 50% in the test-positive cases and is ineffective for the test negatives, and 83.9% if the new treatment reduces hazard 33% uniformly. These are based on a two-sided statistical significance level of 5%. The power of the $z_{\min}$ test does not seem to be superior to that based on a preliminary test of interaction. Although the $z_{\min}$ test is worthy of more detailed study, the interaction approach has the advantage of providing a clearer basis for focusing on either the overall population or the test-positive subset.

Adaptive Designs

Several phase III designs have been proposed that adaptively modify the target patient population that serves as the basis for comparing the new treatment to the control. Wang et al. (55) described a design that is initially planned to accrue N patients of which a fraction, f, are test positive according to a predefined classifier. An interim futility analysis evaluates the effectiveness of the new treatment for the test-negative patients. If the interim analysis indicates that establishing the effectiveness of the new treatment for test-negative patients is futile, then accrual of test-negative patients stops and the final analysis is restricted to evaluating the new treatment in test-positive patients. Otherwise, accrual of test-negative and test-positive patients continues to the target sample size N. At that time, the new treatment is evaluated overall and for test-positive patients. Wang et al. required that a total of N patients be accrued by the end of the trial, even if accrual of the test-negative patients was terminated at the interim analysis. In practice, this could make for a very long clinical trial in the usual case where the proportion of test-positive cases is relatively small. It also makes it difficult to properly compare the performance of their design to other designs that accrue fewer test-positive patients. Early termination of accrual of test-negative patients can be useful in settings where end point evaluation is rapid relative to accrual rate (45).

Biomarker adaptive threshold design. Jiang et al. (56) reported on a "Biomarker Adaptive Threshold Design" for situations where a predictive index is available at the start of the trial but a cutoff point for converting the index to a binary classifier is not established. For example, this design could be used with a FISH assay for EGFR positivity without prespecification of the threshold of positivity (57). With their design, tumor specimens are collected from all patients at entry, but the value of the predictive index is not used as an eligibility criterion. Their analysis plan does not stipulate that the assay for measuring the index needs to be done in real time, although regulators may prefer that the index be used to stratify the randomization between the new treatment and control. Jiang et al. described two analysis plans. Analysis plan A begins with comparing outcomes for all patients receiving the new treatment to those for all control patients. If this difference in outcomes is significant at a prespecified significance level α_1 (e.g., 0.03), then the new treatment is considered effective for the eligible population as a whole. Otherwise, a second-stage test is done using significance threshold $\alpha_2 = 0.05 - \alpha_1$. The second-stage test involves finding the cutoff point b leading to the maximum partial log likelihood for treatment effect $[S(b)]$ when restricted to patients with predictive index at least as large as b. Jiang et al. evaluated the statistical significance of $S(b)$ by randomly permuting the labels of which patients were in the new treatment group and which were controls and determining the maximized partial log likelihood for the permuted data. This is done for thousands of random permutations. If the value $S(b)$ is beyond the $1 - \alpha_2$'th percentile of this null distribution created from the random permutations, then the second-stage test is considered significant.

The advantage of procedure A is its simplicity and that it explicitly separates the test of treatment effect in the broad population from the subset selection. However, the procedure

takes a conservative approach in adjusting for multiplicity of combining the overall and subset tests. An alternative analysis plan B, proposed by Jiang et al., does not use a first-stage comparison of treatment groups overall. Consequently, plan B is more appropriate to settings in which there is greater expectation that treatment effect will be limited to a predictive index–defined subset. With analysis plan B, they determine the cutoff point value b at which $w(b)S(b)$ is maximized, where $w(b)$ is a predefined weight function. They use the weight function to give greater emphasis to the $b = 0$ subset (i.e., the subset containing all patients; predictive index is initially normalized to the 0-1 interval). Let $T(b) = w(b)S(b)$ denote the value of the maximized weighted partial log-likelihood. The statistical significance of $T(b)$ is determined by generating the null distribution by repeating the optimization procedure for many cases of randomly permuted data. With either procedure A or B, a confidence interval for the optimal cutoff point b is generated using a bootstrap resampling procedure. Because the treatment is presumed effective only for patients with predictive index above the threshold b, the confidence coefficient associated with a given value x can be interpreted in a frequentist sense as the probability that a patient with predictive index value x benefits from the new treatment relative to control.

Jiang et al. provided an approach to sample size planning for the biomarker adaptive threshold design. With analysis strategy A, they propose planning sample size in the traditional manner for overall comparison of the treatment arms but powering the trail based on using a significance level α_1 (e.g., 0.03). This involves only a minor increase in sample size compared with the standard approach but provides only limited power for detecting biomarker-restricted treatment effects. With analysis plan B, a larger sample size is used that provides good power for establishing the statistical significance of treatment effects restricted to patients with biomarker values above an initially unknown cutoff point.

Adaptive signature design. For codevelopment of a new drug and companion diagnostic, it is best to have the candidate diagnostic completely specified and analytically validated before its use in the pivotal clinical trials. This is difficult, however, and in some cases is not feasible, particularly with multigene expression–based classifiers. Freidlin and Simon (58) proposed a design for a phase III trial that can be used when no classifier is available at the start of the trial. The design provides for development of the classifier and evaluation of treatment effects in subsets determined by the classifier in a single trial. The analysis plan of the adaptive signature design is structured to preserve the principle of separating the data used for developing a classifier from the data used for evaluating treatment in subsets determined by the classifier, although both processes are part of the same clinical trial.

The analysis plan described by Freidlin and Simon is in two parts as for the design of Simon and Wang (47) described above. At the conclusion of the trial, the new treatment is compared with the control overall using a reduced threshold of significance α_1 (e.g., 0.03). A finding of statistical significance at that level is taken as support of a claim that the treatment is broadly effective. At that point, no biomarkers have been tested on the patients, although patients must have tumor specimens collected to be eligible for the clinical trial.

If the overall treatment effect is not significant at the level α_1, then a second stage of analysis takes place. The patients are divided into a training set and a testing set. Freidlin and Simon used a 50-50 split, but other proportions can be used. The data for patients in the training set are used to define a single subset of patients who are considered most likely to benefit from the new treatment compared with the control. When that subset is explicitly defined, the new treatment is compared with the control for the testing set patients with the characteristics defined by that subset. The comparison of new treatment to control for the subset is restricted to patients in the testing set to preserve the principle of separating the data used to develop a classifier from the data used to test treatment effects in subsets defined by that classifier. The comparison of treatment to control for the subset uses a threshold of significance of α - α_1 (e.g., 0.02) to ensure that the overall chance of a false-positive conclusion does not exceed α (e.g., 0.05).

Friedlin and Simon proposed the adaptive signature design in the context of multivariate gene expression–based classifiers. The size of phase II databases may not be sufficient to develop such classifiers before the initiation of phase III trials (23, 24). Freidlin and Simon showed that the adaptive signature design can be effective for the development and use of gene expression classifiers if there is a very large treatment effect in a subset determined by a set of signature genes. The power of the procedure for identifying the subset is limited, however, by having to test the treatment effect at a very stringent significance level in subset patients restricted to the testing set not used for classifier development.

The analysis strategy used by the adaptive signature design can be used more broadly than in the context of identifying *de novo* gene expression signatures. For example, it could be used when several gene expression signatures are available at the outset and it is not clear which to include in the final statistical testing plan. It could also be used with classifiers based on a single gene but several candidate tests for measuring the expression or deregulation of that gene. For example, the focus may be on EGFR but there may be uncertainty about whether to measure overexpression at the protein level, point mutation of the gene, or amplification of the gene (57). In these settings, with a few candidate classifiers, a smaller training set may suffice instead of the 50-50 split used by Freidlin and Simon.

Conclusions

Developments in cancer genomics and biotechnology are dramatically changing the opportunities for development of more effective cancer therapeutics and molecular diagnostics to guide the use of those drugs. These opportunities can have enormous benefits for patients and for containing health care costs. Achieving these gains, however, requires new approaches to drug development. The current paradigm of post hoc correlative science is not an adequate basis for the

development of predictive oncology. This article has attempted to begin to touch on some alternative approaches for prospective clinical trial design. Codevelopment of drugs and companion diagnostics adds complexity to the already difficult drug development process. Public agencies may have to play a greater role in funding the development of diagnostics for the deregulation of signaling pathways, and regulatory agencies may have to ensure that their policies do not discourage the development of companion diagnostics. To move forward will require focus and a national action plan for removing unnecessary barriers to partnership and progress.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989;8:431–40.
2. Buyse M, Molensberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 2000;1:49–67.
3. Torri V, Simon R, Russek-Cohen E, Midthune D, Freidman M. Relationship of response and survival in advanced ovarian cancer patients treated with chemotherapy. J Natl Cancer Inst 1992;84:407.
4. Chau CH, Rixe O, McLeod H, Figg WD. Validation of analytical methods for biomarkers employed in drug development. Clin Cancer Res. Vol. 18. In press 2008.
5. FDA. Guidance for industry: Pharmacogenomics data submissions. Rockville (MD): Food and Drug Administration, U.S. Department of Health and Human Services; 2005.
6. Pusztai L, Ayers M, Stec J, Hortobagyi GN. Clinical application of cDNA microarrays in oncology. Oncologist 2003;8:252–8.
7. Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 2007;25:5287–310.
8. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–85.
9. Henry NL, Hayes DF. Uses and abuses of tumor markers in the diagnosis, monitoring and treatment of primary and metastatic breast cancer. Oncologist 2006;11:541–52.
10. Hayes DF, Bast RC, Desch CE, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst 1996;88:1456–66.
11. 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.
12. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 2005;97:1180–4.
13. Hayes DF. Prognostic and predictive factors revisited. Breast 2005;14:493–9.
14. Gennari A, Sormani MP, Pronzato P, et al. HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized clinical trials. J Natl Cancer Inst 2008;100:14–20.
15. Hayes DF, Thor AD, Dressler LG, et al. HER2 and response to paclitaxel in node-positive breast cancer. N Engl J Med 2007;357:1496–506.
16. Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer 2007;7:169–81.
17. Toschi L, Cappuzzo F. Understanding the new genetics of responsiveness to epidermal growth factor receptor tyrosine kinase inhibitors. Oncologist 2007;12:211–20.
18. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 2008;26:1626–34.
19. Wolff AC, Hammond EH, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 2007;25:118–45.
20. van't-Veer LJ, Paik S, Hayes DF. Gene expression profiling of breast cancer: a new tumor marker. J Clin Oncol 2005;23:1631–5.
21. Paik S. Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with Tamoxifen. Oncologist 2007;12:631–5.
22. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
23. Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. Biostatistics 2007;8:101–17.
24. Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? Clin Cancer Res 2008;14:108–14.
25. Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. Clin Cancer Res 2007;13:6080–6.
26. West M, Blanchette C, Dressman H. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 2001;98:11462–67.
27. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505–12.
28. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.
29. Simon R, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York: Springer Verlag; 2003.
30. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567–72.
31. Speed TP, editor. Statistical analysis of gene expression microarray data. Chapman and Hall; 2003.
32. Segal MR. Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. Biostatistics 2006;7:268–85.
33. Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. Br J Cancer 2003;89:1599–604.
34. George SL. Statistical issues in translational cancer research. Clin Cancer Res. Vol. 18. In press 2008.
35. Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-ArrayTools. Cancer Informatics 2007;2:11–7.
36. Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. Clin Cancer Res. Vol. 18. In press 2008.
37. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: A comparison of resampling methods. Bioinformatics 2005;21:3301–7.
38. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
39. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147–57.
40. Owzar K, Barry WT, Jung S-H, Sohn I, George SL. Statistical issues arising in the application of genomics and biomarkers in clinical trials. Clin Cancer Res. Vol 18. In press 2008.
41. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 2001;344:783–92.
42. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 2005;10:6759–63.
43. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and correction. Clin Cancer Res 2006;12:3229.
44. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Stat Med 2005;24:329–39.
45. Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol 2004;15:1731–7.
46. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005;23:2020–7.
47. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. Pharmacogenomics J 2006;6:1667–173.
48. Giaccone G, Herbst RS, Manegold C, et al. Gefitinib in combination with gemcitabine and cisplatin in advanced non-small cell lung cancer: a phase III trial-INTACT2. J Clin Oncol 2004;22:777–84.
49. Herbst RS, Giaccone G, Schiller JH, et al. Gefitinib in combination with paclitaxel and carboplatin in advanced non-small-cell lung cancer: a phase III trial-INTACT2. J Clin Oncol 2004;22:785–94.
50. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Branigan BW. Activating mutations in the epidermal growth factor receptor un-

derlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 2004;350:2129–39.

51. Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. Science 2004;304:1497–500.

52. Shepherd FA, Pereira J, Ciuleanu TE, et al. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med 2005;353:123–32.

53. Tsao MS, Sakurada A, Cutz JC, et al. Erlotinib in lung cancer—molecular and clinical predictors of outcome. N Engl J Med 2005;353:133–44.

54. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. Stat Med 2007;26:3535–49.

55. Wang SJ, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat 2007;6:227–44.

56. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst 2007;99:1036–43.

57. Johnson BE, Janne PA. Selecting patients for epidermal growth factor receptor inhibitor treatment: A FISH story or a tale of mutations? J Clin Oncol 2005;28:6813–5.

58. Freidlin B, Simon R. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 2005;11:7872–8.