# Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients

Boris Freidlin and Richard Simon

**Abstract**

**Purpose:** A new generation of molecularly targeted agents is entering the definitive stage of clinical evaluation. Many of these drugs benefit only a subset of treated patients and may be overlooked by the traditional, broad-eligibility approach to randomized clinical trials. Thus, there is a need for development of novel statistical methodology for rapid evaluation of these agents.

**Experimental Design:** We propose a new adaptive design for randomized clinical trials of targeted agents in settings where an assay or signature that identifies sensitive patients is not available at the outset of the study. The design combines prospective development of a gene expression–based classifier to select sensitive patients with a properly powered test for overall effect.

**Results:** Performance of the adaptive design, relative to the more traditional design, is evaluated in a simulation study. It is shown that when the proportion of patients sensitive to the new drug is low, the adaptive design substantially reduces the chance of false rejection of effective new treatments. When the new treatment is broadly effective, the adaptive design has power to detect the overall effect similar to the traditional design. Formulas are provided to determine the situations in which the new design is advantageous.

**Conclusion:** Development of a gene expression–based classifier to identify the subset of sensitive patients can be prospectively incorporated into a randomized phase III design without compromising the ability to detect an overall effect.

Developments in tumor biology have resulted in shift toward molecularly targeted drugs (1–3). Most human tumor types are heterogeneous with regard to molecular pathogenesis, genomic signatures, and phenotypic properties. As a result, only a subset of the patients with a given cancer is likely to benefit from a targeted agent (4). This complicates all stages of clinical development, especially randomized phase III trials (5, 6). In some cases, predictive assays that can accurately identify patients who are likely to benefit from the new therapy have been developed. Then, targeted randomized designs that restrict eligibility to patients with sensitive tumors should be used (7). However, reliable assays to select sensitive patients are often not available (8, 9). Consequently, traditional randomized clinical trails with broad eligibility criteria are routinely used to evaluate such agents. This is generally inefficient and may lead to missing effective agents.

Genomic technologies, such as microarrays and single nucleotide polymorphism genotyping, are powerful tools that hold a great potential for identifying patients who are likely to benefit from a targeted agent (10, 11). However, due to the large number of genes available for analysis, interpretation of these data is complicated. Separation of reliable evidence from the random patterns inherent in high-dimensional data requires specialized statistical methodology that is prospectively incorporated in the trial design. Practical implementation of such designs has been lagging. In particular, analysis of microarray data from phase III randomized studies is usually considered secondary to the primary overall comparison of all eligible patients. Many analyses are not explicitly written into protocols and done retrospectively, mainly as "hypothesis-generating" tools.

We propose a new adaptive design for randomized clinical trials of molecularly targeted agents in settings where an assay or signature that identifies sensitive patients is not available. Our approach includes three components: (*a*) a statistically valid identification, based on the first stage of the trial, of the subset of patients who are most likely to benefit from the new agent; (*b*) a properly powered test of overall treatment effect at the end of the trial using all randomized patients; and (*c*) a test of treatment effect for the subset identified in the first stage, but using only patients randomized in the remainder of the trial. The components are prospectively incorporated into a single phase III randomized clinical trial with the overall false-positive error rate controlled at a prespecified level.

**Authors' Affiliation:** Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland

Received 3/18/05; revised 7/18/05; accepted 8/4/05.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

**Requests for reprints:** Boris Freidlin, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 6130 Executive Boulevard, EPN 8122, MSC 7434, Bethesda, MD 20892-7434. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov.

doi:10.1158/1078-0432.CCR-05-0605

The methodology is presented and evaluated in the context of a binary outcome (e.g., response). With minor adjustment, it can be adapted for use with time-to-event end points, such as survival or disease-free survival.

## Materials and Methods

Consider designing a definitive study to assess whether addition of a new targeted agent to the standard treatment is beneficial. The gold standard for addressing this question is a phase III clinical trial that randomly assigns patients to the combination of the new and standard treatment (arm E) or the standard treatment alone (arm C).

We will describe the proposed design in terms of using DNA microarray expression profiling done to characterize the tumors of the included patients; however, the design is easily adapted to use single nucleotide polymorphism gene typing or proteomic profiling instead. The following modeling assumptions are used: among L evaluated genes, there is a subset of $K$ "sensitivity" genes. The identities of the sensitivity genes are unknown but responsiveness to treatment is influenced by the sensitivity genes through the following model. For the $i$th patient, let $p_i$ denote the probability of response, $t_i$ the treatment that the patient receives ($t_i = 0$ for arm C and $t_i = 1$ for arm E), and $x_{i1}, \ldots, x_{iK}$ the levels of expression for the $K$ unknown sensitivity genes. Then

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \lambda t_i + \gamma_1 t_i x_{\mathrm{il}} + \ldots + \gamma_K t_i x_{iK} \qquad (A)$$

where $\lambda$ is treatment main effect that all patients experience regardless of their gene expression levels and $\gamma_i$ is treatment-expression interaction effect that reflects the degree by which the difference in treatment arms is influenced by the $i$th gene expression level. To simplify the presentation, all gene main effects and the treatment-expression interactions for the nonsensitivity genes are assumed to be 0.

If the interaction variables are positive, patients who overexpress the sensitivity genes have a higher probability of response when treated with the new treatment (E) compared with the standard (C). We assume that a fraction of the patient population overexpresses some (but not necessarily all) of the sensitivity genes. These patients are called "sensitive."

A trial designed to accrue a total of $N$ patients is evaluated in two stages. In stage 1, the first $N_1$ patients are accrued and in stage 2 the remaining $N_2 = N - N_1$ patients are evaluated. A key feature of our design is development of a classifier that predicts whether a patient is more likely to benefit from the new treatment relative to the standard one. This classifier is developed using stage 1 patients only. The classifier is not used to restrict entry of patients during the stage 2 but it is prospectively applied to the stage 2 patients to identify a subset of sensitive patients. The final analysis consists of ($a$) overall comparison of treatment arms E and C using data from all $N = N_1 + N_2$ patients, carried out at significance level $\alpha_1$, and ($b$) comparison of arms E and C in the selected subset of sensitive patients accrued during stage 2, carried out at significance level $\alpha_2$. The study is considered positive if either of the two tests is significant. The overall significance level of this procedure is $\alpha = \alpha_1 + \alpha_2$. Generally, one can use different allocations of the experiment-wise significance level $\alpha$ between the overall effect test and the subset effect test. To preserve the ability of the procedure to detect an overall effect, we recommend setting $\alpha_1$ to 80% of $\alpha$ and $\alpha_2$ to 20% of $\alpha$. For example, setting $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$ corresponds to procedure-wise $\alpha$ level of 0.05. Because the size of the treatment effect in the identified subset may be much greater than in the overall study population, analysis of the subset in patients accrued during the second stage of the trial at a stringent significance level may still provide substantial statistical power.

A large variety of algorithms for developing a classification based on patients accrued during stage 1 could be envisioned. We will describe one such approach here based on machine learning voting methods (12).

Step 1: Using data from stage 1 patients, for each gene $j$ fit the single gene logistic model $\mathrm{logit}(p_i) = \mu + \lambda_j t_i + \beta_j t_i x_{ij}$. Note the genes that have treatment-expression interaction ($\hat{\beta}_j$) significant at a predetermined level $\eta$.

Step 2: Classify stage 2 patients as sensitive or nonsensitive to the new treatment based on the genes with significant interactions in step 1. The $i$th patient in stage 2 is designated sensitive if the predicted new versus control arm odds ratio exceeds a specified threshold ($R$) for at least $G$ of the significant genes $j$ (i.e., $e^{\hat{\lambda}_j + \hat{\beta}_j x_{ij}} > R$).

## Results

We did a simulation study to evaluate performance of the adaptive design. Gene expression levels were generated as follows: ($a$) For sensitivity genes in sensitive patients, using a multivariate normal distribution with mean $m$, variance $\sigma_1^2$, and correlation $\rho$; ($b$) for sensitivity genes in nonsensitive patients, using multivariate normal distribution with mean 0, variance $\sigma_2^2$, and correlation $\rho$; ($c$) for nonsensitivity genes, using multivariate normal with mean 0, variance $\sigma_0^2$, and correlation $\rho$ in all patients. We used $L = 10{,}000$ genes on the array, with the number of sensitivity genes ($K$) either 3, 10, or 20.

Treatment-expression interaction levels were kept constant across sensitivity genes ($\gamma = \gamma_1 = \gamma_2 = , \ldots , = \gamma_K$). For each value of $K$, the interaction levels were scaled to have the same odds ratio $e^5$ between arms E and C for a hypothetical patient with sensitive gene expression levels at their expected value (i.e., $m\gamma K = 5$). We report results for intercept value $\mu$ corresponding to control arm response rate of 25%. Results for other values of control response rates were similar.

To investigate the relationship between gene expression correlation structure and the design performance, we considered two cases: ($a$) an uncorrelated case that assumes, for each patient, that gene expression levels are independent ($\rho = 0$) and ($b$) a highly correlated case that assumes, for each patient, that expression levels of the sensitivity genes are correlated with each other ($\rho = 0.6$) and expression levels of the nonsensitivity genes are correlated with each other ($\rho = 0.6$). In the correlated case for $K = 20$, the sensitivity genes were assumed to come from two independent 10-gene groups with gene expressions correlated within each groups ($\rho = 0.6$). The results are presented in terms of empirical power that is the percentage of the simulated replications of the design that reached the prespecified level of significance. We tabulated empirical powers of the overall arm comparison at 0.05 and 0.04 significance levels and arm comparison in the selected subset at 0.01 significance level. In addition, the overall empirical power of the adaptive design was calculated as the percentage of replications with either positive overall 0.04 level test or positive 0.01 level sensitive subset test.

First, consider a situation where the new treatment effect is restricted to 10% of the eligible patient population that overexpresses 10 sensitivity genes (98% response rate in sensitive patients and 25% response rate in nonsensitive patients on the new treatment arm; 25% response rate for all

control arm patients). A 400-patient trial is carried out. The traditional broad eligibility approach that uses a 0.05 level test has 47% power to detect overall difference between the arms (Fig. 1). In the adaptive approach, overall difference was detected with 43% probability (using a 0.04 level test). Of the 57% of cases where the overall difference was not detected at 0.04 level, the sensitive subset test was significant (at the 0.01 level) in 42%. Thus, the overall power of the adaptive design is 85%, indicating that there is an 85% probability of either detecting a significant overall effect or a significant subset effect. The procedure shows similar ability to identify the subset of sensitive patients in situations where there are 20 or 3 sensitivity genes.

When the gene expressions are correlated, the efficiency of the subset selection is slightly reduced (Fig. 2). When the fraction of sensitive patients is increased to 25%, both the overall 0.05 level test and the adaptive design have over 99% power for detecting the treatment effect (Table 1).

The case where the new treatment effect applies equally to all patients (35% response rate in sensitive and nonsensitive patients on the new treatment arm and 25% response rate on the control arm) is presented in Table 2. The sensitive subset selection algorithm correctly indicates the absence of sensitive subpopulation. At the same time, the overall 0.04 test provides good power for detection of the overall effect. If the new treatment effect is present in both sensitive and nonsensitive patients but the effect is stronger in sensitive patients (99% response rate in sensitive patients and 35% response rate in nonsensitive patients on the new treatment arm; 25% response rate for all control arm patients), the power of the overall test dominates the selected subset test (Table 3). Additional results for a range of model variables are given in Table 4A-D.

## Discussion

The results indicate that development of a gene expression–based classifier to identify the subset of sensitive patients can be prospectively incorporated into a randomized phase III design without compromising the ability to detect an overall effect. Thus, the procedure is especially attractive for allowing pharmaceutical companies to "invest in the development of pharmacogenomic signatures without the risk of losing of broad labeling indications where supported by the results of phase III trials" (13). In addition to a statistically valid procedure for testing for beneficial effect in a subset of patients, the classifier could be instrumental in refining our understanding of the mechanism of action of new agents.

Generally, as the fraction of the sensitive patients increases, so does the difference in overall response rates of arms E and C. Therefore, for fractions above certain threshold, the power of the test for overall effect will dominate the power of the sensitive subset test. The new design preserves the ability to detect the overall effect in this case. However, its advantage over the traditional design (testing for overall effect with broad eligibility) is reduced. In Appendix 1, we present formulas that can assist the investigators in assessing the power relation between the sensitive subset test and the overall test.

In many clinical settings, the total sample size $N$ is fixed by a compromise between considerations of overall effect power

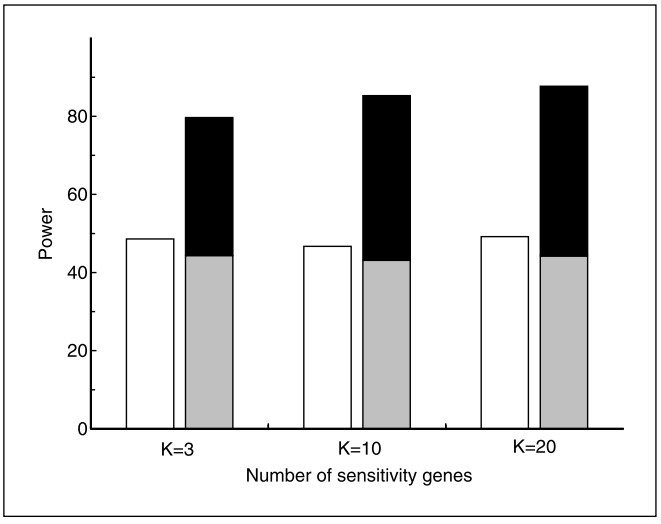


**Fig. 1.** Treatment effect restricted to sensitive patients. Ten percent of patients sensitive. Sensitivity genes are uncorrelated. $N_1 = N_2 = 200$ patients. *Columns,* empirical power (%). Overall 0.05 level test (*white columns*). Overall adaptive signature design (*gray*/*black stacked columns*): overall 0.04 level test (*gray columns*), 0.01 level sensitive subset test but not 0.04 level overall test (*black columns*).

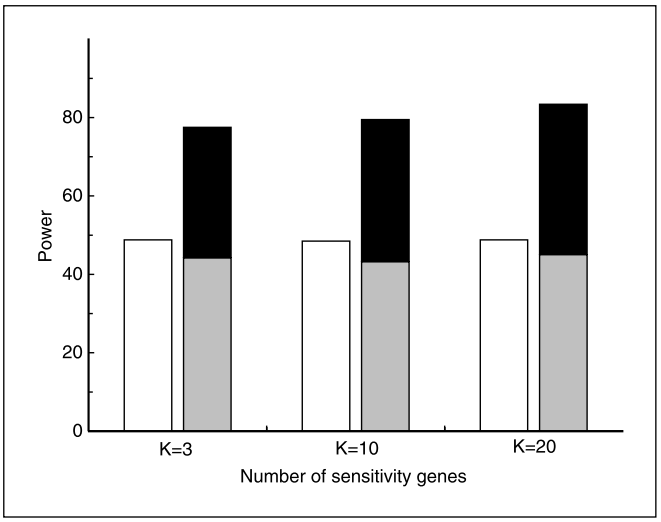


**Fig. 2.** Treatment effect restricted to sensitive patients. Ten percent of patients sensitive. Sensitivity genes are correlated. $N_1 = N_2 = 200$ patients. *Columns,* empirical power (%). Overall 0.05 level test (*white columns*). Overall adaptive signature design (*gray*/*black stacked columns*): overall 0.04 level test (*gray columns*), 0.01 level sensitive subset test but not 0.04 level overall test (*black columns*).

**Table 1.** Empirical power (%)

| Test | Power |
|---|---|
| Overall 0.05 level test* | 99.0 |
| Overall 0.04 level test* | 98.9 |
| Sensitive subset 0.01 level test † | 99.7 |
| Overall adaptive signature design | 99.9 |

NOTE: Treatment effect is restricted to sensitive patients (25% of patients sensitive, 10 sensitivity genes, $N_1 = N_2 = 200$).
*Test based on normal approximation for the difference of two proportions.
†Fisher's exact test.

**Table 2.** Empirical power (%)

| Test | Power |
|---|---|
| Overall 0.05 level test* | 74.2 |
| Overall 0.04 level test* | 70.9 |
| Sensitive subset 0.01 level test † | 1.0 |
| Overall adaptive signature design | 70.9 |

NOTE: Treatment effect applies equally to all patients (10% of patients sensitive, 10 sensitivity genes, $N_1 = N_2 = 200$).
*Test based on normal approximation for the difference of two proportions.
†Fisher's exact test.

and feasibility. For a fixed $N$, the choice of $N_1$ and $N_2$ is based on a tradeoff between the accuracy of the selection procedure that increases with $N_1$ and the size of the stage 2 sensitive patient subset that increases with $N_2$. To preserve the integrity of the design, $N_1$ and $N_2$ need to be defined prospectively. The optimal values of $N_1$ and $N_2$ depend on a number of parameters, including the difference in response between sensitive and nonsensitive patients and the fraction of sensitive patients. Because these are not usually known in advance, we recommend using $N_1 = N_2$. This allocation has been shown to provide robust performance across various settings (see Table 4A- D). It should be noted that the advantage of the adaptive design shown in Figs. 1 and 2 represents a situation where the difference in the new treatment effect between sensitive and nonsensitive patients is large. In settings where this difference is moderate to low, the total sample size ($N = N_1 + N_2$) required to develop and validate the selection procedure may be much larger than needed just for detecting the overall effect.

The optimal values of the tuning parameters $\eta$, $G$, and $R$ depend on the number of sensitive genes $K$, fraction of sensitive patients, and parameters of the logistic model (Eq. A). The true values of the model parameters and the fraction of sensitive patients are not usually known in advance. One can, however, use a cross-validation approach on the stage 1 patients to select tuning parameters values without affecting statistical validity of the procedure. An example of such procedure is provided in Appendix 2.

The issue of selecting the subset of sensitive patients is closely related to the enrichment strategy (14) that uses an intermediate outcome or biomarker to focus on patients that are most likely to benefit from the treatment. Typical enrichment procedures, such as the randomized discontinuation design (15, 16), require a prespecified cutoff value (for the intermediate outcome) to increase the fraction of sensitive patients. Our procedure advances the concept by allowing for prospective selection of a classifier that identifies the patients most likely to benefit from the treatment. Because the classifier is not used to restrict entry of patients to stage 2, its development and application can be carried out at the time of the final analysis. Thus, our procedure can be used with time-to-event end points, such as survival.

A proper implementation of the new design implies using a reduced overall significance level $\alpha_1$ (e.g., 0.04 instead of 0.05) in determining the overall size of the trial. This entails a minor increase in the overall sample size compared with the conventional design (e.g., a 7% increase in sample size for using 0.04 instead of 0.05 significance assuming 90% power). In addition, to avoid bias, sample size for each stage needs to be fixed at the start of the trial.

The gene expression–based classifier, developed on the first-stage patients, is generally quite accurate in situations where the new agent has a strong effect restricted to sensitive patients. In this setting, the new design may substantially reduce the probability of false rejection of effective new treatments. On the other hand, it is important to emphasize that the adaptive

**Table 3.** Empirical power (%)

| Test | Power |
|---|---|
| Overall 0.05 level test* | 97.0 |
| Overall 0.04 level test* | 96.0 |
| Sensitive subset 0.01 level test † | 45.6 |
| Overall adaptive signature design | 97.2 |

NOTE: Stronger treatment effect in sensitive patients (10% of patients sensitive, 10 sensitivity genes, $N_1 = N_2 = 200$).
*Test based on normal approximation for the difference of two proportions.
†Fisher's exact test.

**Table 4A.** Empirical power as a function of total sample size

| Test | Power (%) |
|---|---|
| $N_1 = N_2 = 150$ | |
| Overall 0.05 level | 40.0 |
| Overall 0.04 level | 36.3 |
| Sensitive subset 0.01 level | 37.8 |
| Overall adaptive signature design | 60.5 |
| $N_1 = N_2 = 200$ | |
| Overall 0.05 level | 49.5 |
| Overall 0.04 level | 45.4 |
| Sensitive subset 0.01 level | 75.8 |
| Overall adaptive signature design | 85.7 |
| $N_1 = N_2 = 250$ | |
| Overall 0.05 level | 61.8 |
| Overall 0.04 level | 56.8 |
| Sensitive subset 0.01 level | 93.2 |
| Overall adaptive signature design | 96.2 |
| $N_1 = N_2 = 300$ | |
| Overall 0.05 level | 63.8 |
| Overall 0.04 level | 58.2 |
| Sensitive subset 0.01 level | 98.1 |
| Overall adaptive signature design | 99.0 |
| $N_1 = N_2 = 400$ | |
| Overall 0.05 level | 74.4 |
| Overall 0.04 level | 70.8 |
| Sensitive subset 0.01 level | 99.7 |
| Overall adaptive signature design | 100 |

NOTE: Treatment effect restricted to sensitive patients. Ten percent of patients sensitive, 10 sensitivity genes (98% response rate in sensitive patients and 25% response rate in nonsensitive patients on the new treatment arm; 25% response rate for all control arm patients).

**Table 4B.** Empirical power as a function of proportion of sensitive patients

| Test | Power (%) |
|---|---|
| 5% of patients sensitive | |
| Overall 0.05 level | 19.7 |
| Overall 0.04 level | 17.4 |
| Sensitive subset 0.01 level | 10.3 |
| Overall adaptive signature design | 24.6 |
| 7% of patients sensitive | |
| Overall 0.05 level | 30.4 |
| Overall 0.04 level | 27.4 |
| Sensitive subset 0.01 level | 34.7 |
| Overall adaptive signature design | 50.7 |
| 10% of patients sensitive | |
| Overall 0.05 level | 49.5 |
| Overall 0.04 level | 45.4 |
| Sensitive subset 0.01 level | 75.8 |
| Overall adaptive signature design | 85.7 |
| 15% of patients sensitive | |
| Overall 0.05 level | 77.9 |
| Overall 0.04 level | 74.2 |
| Sensitive subset 0.01 level | 94.4 |
| Overall adaptive signature design | 99.2 |
| 20% of patients sensitive | |
| Overall 0.05 level | 96.2 |
| Overall 0.04 level | 94.7 |
| Sensitive subset 0.01 level | 95.0 |
| Overall adaptive signature design | 100 |
| 25% of patients sensitive | |
| Overall 0.05 level | 99.0 |
| Overall 0.04 level | 98.9 |
| Sensitive subset 0.01 level | 99.7 |
| Overall adaptive signature design | 100 |

NOTE: Treatment effect restricted to sensitive patients. $N_1 = N_2 = 200$, 10 sensitivity genes (98% response rate in sensitive patients and 25% response rate in nonsensitive patients on the new treatment arm; 25% response rate for all control arm patients).

design is protected from violations of the modeling assumptions. Even if the true fraction of sensitive patients is higher than was assumed, or if the selection procedure fails to select the sensitive patients, or if no subset effect is present, the new design provides as much power to detect the overall effect as would be achieved with the standard design. To recapitulate the key aspects of our proposal, consider two phase III trials evaluating a new molecularly targeted agent: (*a*) the traditional broad eligibility design with sample size based on significance level $\alpha$ and (*b*) the adaptive design that has a slightly larger total sample size (based on the reduced significance level $\alpha_1$) with equal number of patents allocated to the first and second stages. If the new agent is beneficial for all patients, both trials have equal power to detect it. At the same time, if the benefit of the new agent is restricted to a subset of the eligible patient population, the second trial may considerably reduce the chance of falsely rejecting the new agent.

There is increasing evidence that patients with the same stage and primary site have tumors that are very different with regard to pathogenesis and the deregulated pathways that are driving tumor growth. Consequently, some molecular targeted agents may only be effective for a small proportion of patients accrued to clinical trials using traditional eligibility criteria. It would be ideal to use the phase II clinical development period for developing an assay or signature for patients most likely to respond to a new agent. For a variety of reasons, however, such biomarkers are often not available by the time phase III trials are initialized. The adaptive design described here may be useful in such situations.

## Appendix 1. Power Estimation

Consider a study with $N/2$ patients per arm and probabilities of response $p_E$ and $p_C$ in arms E and C, respectively. The power of a one-sided $\alpha$ level test to detect a difference in response between the arms is approximately:

$$\text{Power} = \Phi\left\{\frac{(p_E - p_C) - Z_{1-\alpha}\sqrt{\bar{p}(1-\bar{p})\frac{4}{N}}}{\sqrt{p_E(1-p_E)\frac{2}{N} + p_C(1-p_C)\frac{2}{N}}}\right\} \qquad (1.1)$$

where $\bar{p} = (p_E + p_C) / 2$, $\Phi()$ is the normal probability distribution function, and $Z_{1-\alpha}$ is $(1-\alpha)$th percentile of normal distribution.

**Table 4C.** Empirical power as a function of $N_1/N_2$ ratio

| Test | Power (%) |
|---|---|
| $N_1 = 150$, $N_2 = 250$ | |
| Overall 0.05 level | 49.4 |
| Overall 0.04 level | 45.0 |
| Sensitive subset 0.01 level | 72.7 |
| Overall adaptive signature design | 86.4 |
| $N_1 = 174$, $N_2 = 226$ | |
| Overall 0.05 level | 49.6 |
| Overall 0.04 level | 44.7 |
| Sensitive subset 0.01 level | 76.4 |
| Overall adaptive signature design | 87.7 |
| $N_1 = 200$, $N_2 = 200$ | |
| Overall 0.05 level | 49.5 |
| Overall 0.04 level | 45.4 |
| Sensitive subset 0.01 level | 75.8 |
| Overall adaptive signature design | 85.7 |
| $N_1 = 226$, $N_2 = 174$ | |
| Overall 0.05 level | 48.7 |
| Overall 0.04 level | 43.9 |
| Sensitive subset 0.01 level | 51.2 |
| Overall adaptive signature design | 69.2 |
| $N_1 = 250$, $N_2 = 150$ | |
| Overall 0.05 level | 50.0 |
| Overall 0.04 level | 46.7 |
| Sensitive subset 0.01 level | 52.5 |
| Overall adaptive signature design | 71.9 |

NOTE: Treatment effect restricted to sensitive patients. $N_1 + N_2 = 400$, 10 sensitivity genes (98% response rate in sensitive patients and 25% response rate in nonsensitive patients on the new treatment arm; 25% response rate for all control arm patients).

**Table 4D.** Empirical power as a function of difference in response rates between sensitive and nonsensitive patients on the new treatment arm

| Test | Power (%) |
|---|---|
| 98% response rate in sensitive patients and 25% response rate in nonsensitive patients ($\gamma = 0.5$) | |
| Overall 0.05 level | 49.5 |
| Overall 0.04 level | 45.4 |
| Sensitive subset 0.01 level | 75.8 |
| Overall adaptive signature design | 85.7 |
| 95% response rate in sensitive patients and 25% response rate in nonsensitive patients ($\gamma = 0.4$) | |
| Overall 0.05 level | 43.0 |
| Overall 0.04 level | 38.5 |
| Sensitive subset 0.01 level | 63.1 |
| Overall adaptive signature design | 75.0 |
| 87% response rate in sensitive patients and 25% response rate in nonsensitive patients ($\gamma = 0.3$) | |
| Overall 0.05 level | 36.7 |
| Overall 0.04 level | 31.7 |
| Sensitive subset 0.01 level | 34.5 |
| Overall adaptive signature design | 51.6 |
| 80% response rate in sensitive patients and 25% response rate in nonsensitive patients ($\gamma = 0.25$) | |
| Overall 0.05 level | 31.6 |
| Overall 0.04 level | 28.4 |
| Sensitive subset 0.01 level | 17.6 |
| Overall adaptive signature design | 38.8 |
| 71% response rate in sensitive patients and 25% response rate in nonsensitive patients ($\gamma = 0.2$) | |
| Overall 0.05 level | 26.0 |
| Overall 0.04 level | 22.6 |
| Sensitive subset 0.01 level | 6.3 |
| Overall adaptive signature design | 26.3 |

NOTE: Treatment effect restricted to sensitive patients. $N_1 = N_2 = 200$, 10 sensitivity genes (25% response rate for all control arm patients).

The expected probability of response for a patient receiving treatment E can be written in terms of the model variables as:

$$p_E = F_S \frac{e^{\mu+\lambda+K\gamma m}}{1+e^{\mu+\lambda+K\gamma m}} + (1-F_S)\frac{e^{\mu+\lambda}}{1+e^{\mu+\lambda}} \quad (1.2)$$

where $F_S$ denotes the fraction of sensitive patients. For the control arm C, the expected response probability is

$$p_C = \frac{e^{\mu}}{1+e^{\mu}}. \quad (1.3)$$

The power of the overall arm comparison is obtained by substituting Eqs. 1.2 and 1.3 in Eq. 1.1.

For a sensitive patient on arm E, the expected response probability is:

$$p_{S_E} = \frac{e^{\mu+\lambda+K\gamma m}}{1+e^{\mu+\lambda+K\gamma m}} \quad (1.4)$$

The subset selection algorithm is usually subject to some error. Let $p_{sens}$ denote the sensitivity of the subset selection algorithm (the probability that a sensitive patient is selected) and $p_{spec}$ denote the specificity (the probability that a nonsensitive patient is not selected). The probability that a selected patient is sensitive, called the positive predictive value (PPV), is

$$\mathrm{PPV} = \frac{F_S p_{sens}}{F_S P_{sens} + (1-F_S)(1-p_{spec})}.$$

The expected response probability for a patient receiving treatment E in the selected subset is

$$p_{+E} = \mathrm{PPV} p_{S_E} + (1-\mathrm{PPV})\frac{e^{\mu+\lambda}}{1+e^{\mu+\lambda}}$$

The expected response probability for a patient receiving treatment C in the selected subset is $P_{+C} = P_C$. The expected size of the selected subset is $N_+ = N_2\,[F_S\,p_{sens} + (1 - F_S)(1 - p_{spec})]$. Therefore, the power of the subset comparison is obtained by substituting $N_+$ for $N$ into Eq. 1.1 and using $p_{+E}$ instead of $p_E$.

For the design purposes, we recommend evaluation of the adaptive design for $p_{sens}$ and $p_{spec}$ values in the range of 0.9 to 1.0.

## Appendix 2. Selection of Tuning Parameters

In the simulation study, the tuning parameters were selected empirically by choosing the values that gave the highest power on a separate set of replications. In practice, we recommend the following approach based on leave-one-out cross-validation to select the best combination of $\eta$, $G$, and $R$ from a set of $M$ possible combinations (using stage 1 patients only):

Part 1: Remove the $i$th patient and carry out step 1 of the two-step subset selection procedure (described in Materials and Methods) on the remaining $(N_1 - 1)$ patients. Then, using step 2 of the two-step procedure, determine if the left-out patient is classified as sensitive according to each of the $M$ possible tuning parameter combinations.

Part 2: Repeat part 1 on each stage 1 patient and form $M$ subsets of sensitive patients, each corresponding to a set of tuning parameters.

Part 3: Compare arms E and C in each of the $M$ subsets. Select the tuning variable combination that provides to the smallest $P$ value in comparing treatments. This approach preserves the validity of the subset selection procedure as only the data from the first phase is used to determine the tuning parameters.

## Acknowledgments

We thank the referees for their valuable comments.

## References


**1.** Balis FM. Evolution of anticancer drug discovery and the role of cell based screening. J Natl Cancer Inst 2002;94:78–9.

**2.** Schilsky RL. End points in cancer clinical trials and the drug approval process. Clin Cancer Res 2002;8: 935–8.

**3.** Rothenberg ML, Carbone DP, Johnson DH. Improving the evaluation of new cancer treatments: challenges and opportunities. Nat Rev Cancer 2003;3: 303–9.

**4.** Grunwald V, Hidalgo M. developing inhibitors of the epidermal growth factor receptor for cancer treatment. J Natl Cancer Inst 2003;95:851–67.

**5.** Betensky RA, Louis DN, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomized clinical trials. J Clin Oncol 2002;20:2495–9.

**6.** Freidlin B, Korn E. A testing procedure for survival data with few responders. Stat Med 2004;23: 1818–23.

**7.** Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trails. Clin Cancer Res 2004;10:6759–63.

**8.** Dancey JE, Freidlin B. Targeting epidermal growth factor receptor—are we missing the mark? Lancet 2003; 362:62–6.

**9.** Johnson DH. Targeted therapy in non-small cell lung cancer: myth or reality. Lung Cancer 2003;41:s3–8.

**10.** Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nat Med 2002;8:68–74.

**11.** Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. N Engl J Med 2002;346:1937–47.

**12.** Breiman L. Bagging predictors. Mach Learn 1996; 24:123–40.

**13.** Simon R. An agenda for clinical trials in the genomic era. Clin Trials 2004;1:468–70.

**14.** Temple RJ. Special study designs: early escape, enrichment, studies in non-responders. Commun Stat Theory Methods 1994:23:499–531.

**15.** Stadler WM, Ratain MJ. Development of target-based antineoplastic agents. Invest New Drugs 2000; 18:7–16.

**16.** Freidlin B, Simon R. An evaluation of randomized discontinuation design. J Clin Oncol 2005;23: 5094–8.