Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect

Wenyu Jiang, Boris Freidlin, Richard Simon

Background Many molecularly targeted anticancer agents entering the definitive stage of clinical development benefit only a subset of treated patients. This may lead to missing effective agents by the traditional broad-eligibility randomized trials due to the dilution of the overall treatment effect. We propose a statistically rigorous biomarker-adaptive threshold phase III design for settings in which a putative biomarker to identify patients who are sensitive to the new agent is measured on a continuous or graded scale.

Methods The design combines a test for overall treatment effect in all randomly assigned patients with the establishment and validation of a cut point for a prespecified biomarker of the sensitive subpopulation. The performance of the biomarker-adaptive design, relative to a traditional design that ignores the biomarker, was evaluated in a simulation study. The biomarker-adaptive design was also used to analyze data from a prostate cancer trial.

Results In the simulation study, the biomarker-adaptive design preserved the power to detect the overall effect when the new treatment is broadly effective. When the proportion of sensitive patients as identified by the biomarker is low, the proposed design provided a substantial improvement in efficiency compared with the traditional trial design. Recommendations for sample size planning and implementation of the biomarker-adaptive design are provided.

Conclusions A statistically valid test for a biomarker-defined subset effect can be prospectively incorporated into a randomized phase III design without compromising the ability to detect an overall effect if the intervention is beneficial in a broad population.

J Natl Cancer Inst 2007;99:1036–43

Human cancers are heterogeneous with regard to their molecular and genomic properties. Recent advances in biotechnology have resulted in a shift toward molecularly targeted anticancer agents. These new therapeutics are likely to benefit only a subset of the patients with a given cancer. Definitive testing of such targeted agents requires the identification of the appropriate "sensitive" population. When biomarkers to identify the patients who are likely to benefit from the new therapy are available, targeted clinical trials that restrict eligibility to sensitive patients should be used (1). However, reliable assays to identify sensitive patients are often unavailable. In the absence of a reliable biomarker, broad-eligibility clinical trials are used routinely. Most of these trials use a conventional design, in which the primary analysis is based on comparison of all randomly assigned patients. This often leads to the failure to recognize effective agents due to dilution of the treatment effect by the presence of the patients who do not benefit from the agent. Retrospective analysis of trials with a conventional design can be used as an initial step in identifying biomarkers for the sensitive subpopulation. However, retrospectively identified biomarkers typically have to be validated in a confirmatory prospective randomized phase III clinical trial (2). This approach is inefficient and may considerably prolong clinical development.

Previously, we have proposed a design [adaptive signature design (3)] that combines a definitive test for treatment effect in a broad population with identification and validation of a genomic signature for the subset of sensitive patients if the broad population test is negative. The adaptive signature design was developed for high-dimensional data such as gene expression microarrays, where only a few unknown genes among thousands assayed may be relevant and where a classifier (signature) to identify sensitive patients is not available. The design incorporates both the identification and the validation of a pharmacogenomic signature for sensitive patients.

Often, preliminary information on a biomarker to identify the sensitive subset of patients is available but an appropriate cutoff

Affiliation of authors: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD.

Correspondence to: Boris Freidlin, PhD, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, EPN-8122, National Cancer Institute, Bethesda, MD 20892 (e-mail: freidlinb@ctep.nci.nih.gov).

See "Notes" following "References."

DOI: 10.1093/jnci/djm022

value to separate sensitive from insensitive patients has not been properly established (or validated). Many biomarkers that are originally measured on a continuous or graded scale (such as the proportion of cells in S phase) are then used to categorize patients into several distinct categories for clinical management. In particular, pharmacogenomically developed drugs often rely on assays to measure target expression levels (e.g., HER2 or epidermal growth factor receptor [EGFR]) on a graded scale; these levels are then converted to dichotomous (positive/negative) status based on a cutoff that has not been properly validated. For example, various cut points have been used when EGFR is measured immunohistochemically with the DAKO kit, and the "optimal" threshold has not been established (4).

Here, we propose a statistically rigorous phase III trial design for settings in which a putative biomarker is measured on a continuous or graded scale. The design combines a test for overall treatment effect in all randomly assigned patients with the establishment and validation of a cut point for a prespecified biomarker for identifying the sensitive subpopulation. The procedure provides prospective tests of the hypotheses that the new treatment is beneficial for the entire patient population or that it is beneficial for a subset of patients defined by the biomarker.

In addition to a formal test for any treatment effect (overall or in a subpopulation), our procedure provides an estimate of the optimal biomarker cutoff and a graph for estimating the probability that a patient with given biomarker value would benefit from the new therapy. These tools can be useful in selecting a biomarker-defined subpopulation with an improved risk–benefit ratio as well as in guiding treatment decisions for individual patients.

CONTEXT AND CAVEATS

Prior knowledge

Many molecularly targeted anticancer agents have the potential to benefit only a subset of patients, that is, those whose levels of the target exceed a certain threshold level. When a biomarker for the target is available but a cutoff to distinguish sensitive from insensitive patients has not been defined, a clinical trial will include insensitive as well as sensitive patients, and any effect of the agent on the subset of sensitive patients may therefore be missed.

Study design

A phase III trial design was developed that combines a test for treatment effect in all patients with the identification and validation of a cutoff point for a prospectively chosen biomarker. The design was tested in a simulation study and was also used to analyze data from an existing trial.

Contribution

In the simulation study, the design allowed a benefit to be seen both when the agent was effective in a broad patient population and when it was effective in a smaller, biomarker-defined subset. For example, when the treatment causes a 79% reduction in hazard in just 10% of the patients, the trial design has a power of 63% whereas a standard design would have a power of only 24%.

Implications

Using this design, it should be possible to prospectively incorporate validation of a biomarker for identifying sensitive patients into a randomized phase III trial design in such a way that an overall effect can still be detected if one exists.

Limitations

The approach has not yet been tested in an actual clinical trial. Some increase in sample size may be necessary. The approach requires that a quantitative biomarker for sensitivity has already been identified.

Methods

Design Considerations

The gold standard for definitive evaluation of a new agent is a randomized phase III trial. Consider a trial that is designed to assess whether the addition of a new therapy to standard care is beneficial. In such a trial, patients are randomly assigned to receive the combination of the new and standard treatment (experimental arm) or the standard treatment alone (control arm). The two treatment arms are compared with respect to time (t) to a clinical event, such as death or disease progression. Time-to-event data of this type are frequently modeled using the proportional hazards model (5). The hazard function denotes the instantaneous risk of the event (e.g., death or disease progression) as a function of time t. When the data conform to the proportional hazards assumption, the logarithm of the ratio of the hazard function for patients in the experimental treatment arm to the hazard function for patients in the control arm is a constant independent of time. This model is conventionally written as

$$\log\left(h_E(t)/h_C(t)\right) = \gamma, \qquad [1]$$

where $h_E(t)$ and $h_C(t)$ denote the hazard functions for experimental and control arms, respectively, and γ denotes treatment effect. In a standard broad-eligibility trial, the primary analysis is often based on the log likelihood ratio statistic (S) for overall treatment effect. A log likelihood ratio statistic is simply the minus log of the ratio of the likelihood of the data under the null hypothesis that there is no treatment benefit to the likelihood of the data without such restriction.

Now, consider a setting for which there is preliminary evidence that the new therapy may be beneficial in only a "sensitive" subset of patients, as defined by a quantitative biomarker, and that the risk–benefit ratio will be maximized in patients with biomarker levels above some unknown cutoff value. We formalize this in the following model:

$$\log\left(h_E(t)/h_C(t)\right) = \begin{cases} 0 & \text{for patients with biomarker values below } c_0 \\ \gamma & \text{for patients with biomarker values above } c_0 \end{cases}, \qquad [2]$$

where c_0 denotes the unknown cutoff value. This model assumes that patients with biomarker values above c_0 benefit from the new treatment and that patients with biomarker values below c_0 do not. Without loss of generality, it is assumed that the biomarker (and c_0) take values in the interval (0, 1). Model [2] reduces to model [1] when $c_0 = 0$.

Model [2] is sometimes referred to as a "cut-point" model (6). It is a simplified version of a more general model that describes the relationship among outcome, treatment, and biomarker value. For a technical discussion of the models, see Appendix A.

Procedures A and B

The objective of the design is to determine whether 1) the experimental arm is better than the control arm for all randomly assigned patients, 2) the experimental arm is better than the control arm for a subset of patients defined by values of biomarker greater than some value c_0, or 3) the experimental arm is not better than the control arm. The adaptive signature design (3) makes it possible to combine development of a pharmacogenomic signature for selecting a subset of sensitive patients with a statistically rigorous integration of both a test of the new treatment overall (for all randomly assigned patients) and a test of the new treatment in the adaptively defined subset. The same strategy for combining two tests can be applied in the present setting, as follows. First, a test for treatment effect in all patients is conducted at a reduced significance level, α_1. If the test is statistically significant, then for the purpose of formal evaluation, the procedure stopped and the null hypothesis of no treatment effect for the randomized patients as a whole is rejected. Otherwise, an algorithm (see "Simulation Study" below) is applied to test for treatment effect in a biomarker-defined subset of patients at a significance level of $\alpha_2 = \alpha - \alpha_1$ (where α is typically .05). This procedure, hereafter referred to as procedure A, controls the probability of making any false-positive claim at the prespecified level α. To preserve the ability of procedure A to detect an overall effect, we recommend setting α_1 to 80% of α and α_2 to 20% of α (see Appendix A for a discussion on the choice of α_1 and α_2). For example, setting $\alpha_1 = .04$ and $\alpha_2 = .01$ corresponds to a procedure-wise α level of .05.

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. Thus, at the end of the study, one of the three distinct outcomes is established: 1) treatment benefit is shown in a broad population, 2) treatment effect is shown in a biomarker-defined subset of patients, or 3) no treatment effect is detected. However, the procedure takes a conservative approach in adjusting for multiplicity that results from combining the overall and subset tests.

Procedure B is a generalization of procedure A that is based on a more efficient approach to combining the overall and subset tests by incorporating the correlation structure of the test statistics. For each candidate biomarker cutoff value in the interval (0, 1), model [2] is fitted on the subset of patients with biomarker values above that cutoff value and a log likelihood ratio statistic is calculated for testing the hypothesis that there is no treatment effect in patients with biomarker exceeding that value (note that for a cutoff value of 0, the likelihood ratio statistic is the overall effect statistic, S). A natural approach to converting a series of statistics that are calculated over the range of possible cutoff values into a single test is to take the maximum (7). To ensure that the resulting procedure has reasonable power when the new treatment is effective for the entire population, we weight up the contribution of the overall test S. This is achieved by adding a positive constant R to statistic S before taking the maximum; the resulting statistic is denoted T. Adding a constant to S is a generalization of the approach taken in procedure A, where a higher portion of the procedure-wise error rate is allocated to the overall effect test. We recommend $R = 2.2$, which we found to provide the best balance between the ability of procedure B to detect an overall effect and its ability to detect a subset effect (see Appendix A).

Statistic T uses a cutoff value that is optimized over the range of possible cutoff points. Because of the well-known multiple testing problem, the standard asymptotic theory does not apply (6,8). To provide a statistically valid P value, we use the permutation distribution of the test statistic T, in which the treatment group labels are permuted (9); see Appendix A for details.

If procedure B rejects the null hypothesis of no treatment effect, the next step is to identify the biomarker threshold above which the new treatment is more effective than the control. We obtain both a point estimate and a confidence interval (CI) for the cutoff c_0, as described in Appendix A. In addition, we propose using a graphical representation of the distribution function of the estimate of c_0 as a convenient tool for communicating the study result to patients and clinicians. In situations where there is no overall treatment effect, this graph is interpreted as the probability that a patient with a given biomarker value will benefit from the new treatment. It can therefore be used as a clinical management tool for guiding individual patient decisions. In theory, one can obtain the estimate of cutoff c_0 even if procedure B did not reject the null hypothesis. However, we do not recommend doing so because the estimate will be difficult to interpret.

It is important to prospectively incorporate the biomarker-based procedures into the study design and describe them in the study protocol. In particular, the procedures require careful sample size planning (see "Sample Size Considerations" below).

Simulation Study

We conducted a simulation study to evaluate the performance of procedures A and B relative to a standard broad eligibility phase III design (based on testing for overall effect in the entire study population). In procedure A, the second-stage biomarker-defined subset effect test was based on the permutation distribution of the maximized log likelihood ratio statistic with cutoff value range restricted to the interval (0.5, 1). (This creates a test focused on treatment effect that is limited to a relatively small subpopulation and is therefore unlikely to be detected by the overall effect test.)

The simulations corresponded to clinical trials with 200 patients randomly assigned between the experimental and control arms. Outcome data were generated from an exponential model ($h_E(t) = h_C(t) = 1$). Biomarker values were generated from a uniform distribution on the interval (0, 1). Administrative censoring resulting from staggered entry ranged between 10% and 20%. We used a grid of candidate biomarker cutoff values of 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Procedure A used $\alpha_1 = .04$ and $\alpha_2 = .01$. Constant $R = 2.2$ was used in procedure B.

For the first set of simulations, it was assumed that all patients benefit from the new therapy (corresponding to model [1]). A second set of simulations generated data corresponding to model [2], in which only patients with biomarker values above a certain cutoff value benefit from the new treatment. We performed simulations with cutoff values of $c_0 = 0.25$, 0.5, 0.75, and 0.9. To evaluate the performance of the proposed procedures under a departure from model [2], we considered two additional situations. In one ("linear trend"), the log hazard ratio (HR) of benefit of the experimental over the standard arm increases linearly over the entire range of biomarker values. In the other ("delayed linear trend"), there was no benefit (log HR = 0) for patients with biomarker value below

0.5 and a linear increase in log hazard ratio for patients with biomarker values above 0.5.

The simulation 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 statistical significance.

Results

Simulations

Simulation results for the overall treatment effect test and for procedures A and B are presented in Table 1. For each setting, the different rows represent different magnitudes of the treatment effect.

Simulation 1 addresses the situation in which the new therapy is beneficial for all patients. Not surprisingly, the overall effect test had the highest power under this model. However, procedures A and B had similar power, with only marginal loss relative to that of the overall test. Simulations 2–5 address situations in which only a proportion of patients benefit from the new therapy. The standard phase III design, with its reliance on the test for overall effect, generally resulted in a considerable loss of power. Indeed, the simulations showed that the ability of the standard design to detect treatment effect decreased as the proportion of the sensitive patients in the population decreases. For example, for a 43% reduction in hazard (in sensitive patients), the power to detect the benefit in the overall test was reduced from 97% when all patients benefit to less than 20% when only 25% of patients (those with biomarker values greater than 0.75) benefit. The proposed procedures provide improved power under the subset effect scenarios 2 through 5. In many instances, the improvement is substantial. For example, when there is a 79% reduction in hazard in 10% of patients (only in those with biomarker values above 0.90; simulation 5), the power of procedure A is 63% compared with 24% for the overall test. Under a linear trend model (simulation 6), the proposed procedures had slightly better power than the standard design. Under the delayed trend model, the power advantage of procedure B was considerable (simulation 7).

Thus, the simulation study showed that procedures A and B provide good power for detecting a subset effect while preserving the ability to detect an overall treatment effect when it is present. In general, procedure B has a higher power than procedure A across the settings considered.

Application to Prostate Cancer Data

To illustrate how the proposed procedures would work in practice, we used data from the second Veterans Administration Cooperative Urologic Research Group clinical trial (10,11). This double-blind clinical trial randomly allocated 506 prostate cancer patients to one of four arms: placebo, 0.2 mg of diethylstilbestrol (DES), 1.0 mg DES, or 5.0 mg DES. Similar to Byar and Corle (10), in our analysis, the two lower doses (placebo and 0.2 mg DES) were combined in a single arm (designated arm C) and the two higher doses (1.0 mg DES and 5.0 mg DES) were combined in a single arm (designated arm E). Arms E and C were then compared with respect to overall survival (i.e., death from any cause).

Table 1. Empirical power of procedures A and B versus the overall test*

Simulation	Model	% Reduction in hazard (hazard ratio)	Empirical power†		
			Overall test	Procedure A	Procedure B
1	Everybody benefits from new therapy	20 (0.8)	.330	.304	.313
		33 (0.67)	.775	.751	.732
		43 (0.57)	.965	.957	.943
2	Only patients with biomarker values > 0.25 benefit from new therapy	43 (0.57)	.819	.802	.837
		60 (0.4)	.996	.997	.998
3	Only patients with biomarker values > 0.5 benefit from new therapy	43 (0.57)	.505	.562	.607
		60 (0.4)	.888	.932	.952
4	Only patients with biomarker values > 0.75 benefit from new therapy	43 (0.57)	.196	.280	.311
		60 (0.4)	.429	.604	.641
		69 (0.31)	.600	.806	.846
5	Only patients with biomarker values > 0.9 benefit from new therapy	60 (0.4)	.105	.238	.274
		69 (0.31)	.162	.401	.412
		79 (0.21)	.238	.632	.624
6	Linear increase in hazard ratio	43 (0.57)‡	.497	.504	.542
		60 (0.4)‡	.887	.892	.909
		69 (0.31)‡	.974	.981	.985
7	Linear increase in hazard ratio for patients with biomarker values > 0.5	43 (0.57)§	.166	.212	.262
		60 (0.4)§	.386	.514	.541
		69 (0.31)§	.559	.744	.741

* For each model, data were generated with 2–3 hazard reduction levels.

† Empirical power is the percentage of the simulated replications of the design that reached the prespecified level of statistical significance (a two-sided .05 significance level).

‡ Maximum effect, effect increases linearly in biomarker from 0 to the maximum.

§ Maximum effect, effect increases linearly in biomarker from 0.5 to the maximum.

Table 2. Second Veterans Administration Cooperative Urologic Research Group clinical trial data*

Variable	No. of patients with measured covariate	*P* value		
		Overall test	Procedure A, Stage 2†	Procedure B‡
AP	505	.084	.019	.041
SG	494	.110	.025	.050

* AP = prostatic acid phosphatase level; SG = combined index of tumor stage and histologic grade.

† Based on the permutation distribution of the maximized log likelihood ratio statistic with cutoff value range restricted to interval (0.5, 1).

‡ Procedure B used $R = 2.2$.

We investigated two biomarkers in the analysis of the prostate cancer data, serum prostatic acid phosphatase (AP) and the combined index of tumor stage and histologic grade (SG), to determine whether either biomarker can be used to identify a subset of patients for whom DES is beneficial. The variable AP measures the serum prostatic AP level in King-Armstrong units; the numbers are continuous from 1 to 5960, with a median value of 7. The variable SG records the combined index of tumor stage and histologic grade; it takes integer values ranging from 5 to 15, with a median of 10.

As Table 2 shows, a standard clinical trial design based on testing overall treatment effect (in all randomly assigned patients with nonmissing AP values) failed to detect a benefit for DES at the .05 significance level ($P = .084$). We next applied procedure A, first performing the overall test at a .04 significance level. Because this significance level was not reached, we proceeded to testing for subset effect at the .01 level. As Table 2 indicates, the subset test *P* value was .019. Thus, procedure A failed to reject the null hypothesis of no effect (at an overall .05 level). Procedure B yielded a *P* value of .041, indicating that treatment was effective in a subset of the population with high AP values. The estimated cutoff value (Table 3) was 36 (95% CI = 9 to 170). This cutoff corresponded to a 77th percentile of the study population AP values. The estimated probability of benefit for a patient with a given value of AP is summarized in Fig. 1. For example, a patient with AP value of 100 is estimated to have more than 90% probability of benefit, whereas a patient with AP value of 20 has only about 10% chance of benefit.

Similarly, for the SG biomarker, procedure A failed to reject the null hypothesis (overall test $P = .11$, subset effect test $P = .025$). Procedure B indicated that DES is beneficial in a subset of patients with high SG values ($P = .05$). The cutoff value was estimated to be 11 (95% CI = 10 to 13; Table 3)—that is, patients with SG values

Table 3. Second Veterans Administration Cooperative Urologic Research Group clinical trial data: cutoff estimate and confidence intervals*

Variable	No. of patients with measured covariate	Estimated cutoff level (95% CI)
AP	505	36 (9 to 170)
SG	494	11 (10 to 13)

* AP = prostatic acid phosphatase level; SG = combined index of tumor stage and histologic grade; CI = confidence interval.

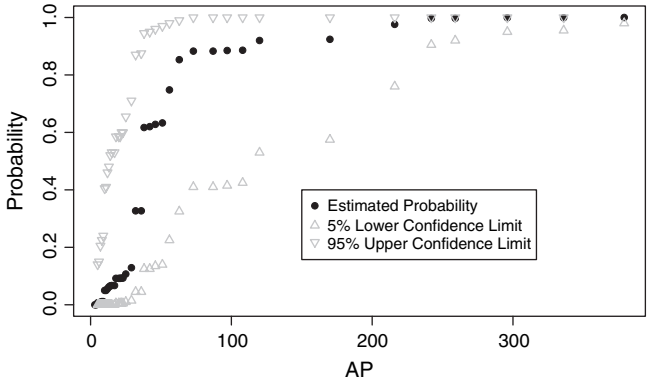

Fig. 1. Estimated probability that a prostate cancer patient with a given value of serum prostatic acid phosphatase (AP) will benefit from treatment with diethylstilbestrol. Data come from the second Veterans Administration Cooperative Urologic Research Group clinical trial (10,11). AP level is given in King-Armstrong units.

above 11 (corresponding to the 53rd percentile of the study population SG distribution) benefited from the treatment. The estimated probability of benefit for a patient with a given value of SG is summarized in Fig. 2. For example, a patient with SG value of 12 is estimated to have more than 90% probability of benefit, whereas a patient with SG value of 10 has less than 5% probability of benefit.

Sample Size Considerations

The simulation results in Table 1 were then used to develop an approach to sample size planning for procedures A and B. Many new agents coming to the definitive stage of clinical testing have been developed to focus on a specific biologic target and thus may benefit only a subset of patients with a given disease. For practical considerations, however, most definitive clinical trials are designed with eligibility criteria that focus on the broadest possible population in which beneficial effect can be reasonably expected. Therefore, the sample size calculations are based on having adequate power for detecting an overall treatment effect in all randomly assigned patients. The target treatment effect is typically set at the minimal clinically meaningful benefit (for example, many trials in cancer are designed to detect a 25% reduction in hazard).

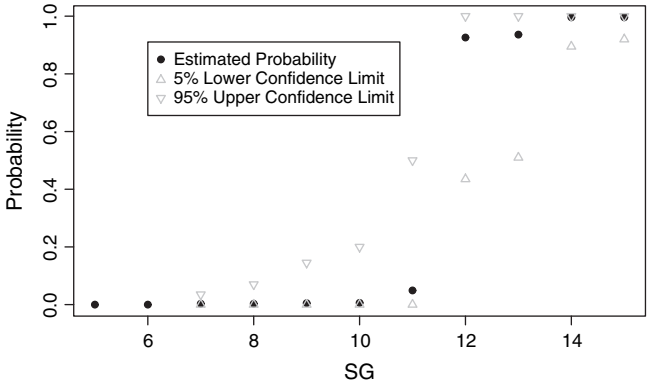

Fig. 2. Estimated probability that a prostate cancer patient with a given value on a combined index of tumor stage and histologic grade (SG) will benefit from treatment with diethylstilbestrol. Data come from the second Veterans Administration Cooperative Urologic Research Group clinical trial (10,11). The SG variable takes integer values of from 5 to 15.

Table 4. Estimated power of broad eligibility design versus biomarker-adaptive design A*

Model	Estimated power, %	
	Broad eligibility design	Biomarker-adaptive design A†
40% reduction in hazard in 50% of patients defined by the biomarker (22% reduction in overall hazard)	70	78
60% reduction in hazard in the 25% of patients defined by the biomarker (20% reduction in overall hazard)	65	91
79% reduction in hazard in 10% of patients defined by the biomarker (14% reduction in overall hazard)	35	93

* Based on a trial with approximately 406 events and a two-sided .05 statistical significance level.

† Based on relative efficiency from Table 1.

Table 5. Approximate number of events required to detect a 60% reduction in hazard in a subset of patients using broad eligibility design versus biomarker-adaptive design B

Model	No. of events*	
	Broad eligibility design	Biomarker-adaptive design B†
Only patients with biomarker values above 0.25 benefit from new therapy (75% of patients are sensitive)	68	68
Only patients with biomarker values above 0.5 benefit from new therapy (50% of patients are sensitive)	150	132
Only patients with biomarker values above 0.75 benefit from new therapy (25% of patients are sensitive)	600	316
Only patients with biomarker values above 0.9 benefit from new therapy (10% of patients are sensitive)	3740	998

* For 80% power at two-sided .05 significance level.

† Numbers are based on relative efficiency from Table 1 for a hazard ratio of 0.4.

A simple approach to sample size planning is to use the standard broad-eligibility sample size calculation with a reduced statistical significance level (e.g., using the .04 significance level instead of .05) and to follow the analysis plan of procedure A (as defined in "Methods"). This approach (biomarker-adaptive design A) ensures no reduction in statistical power for the overall analysis, provides an opportunity to establish the effectiveness of the new treatment for a subset when the overall test is negative, and entails only a modest increase in sample size compared with a trial that ignored the biomarker.

Consider a typical phase III clinical trial in cancer designed with 80% power to detect a 25% reduction in hazard (at a two-sided .05 significance level). This standard broad-eligibility design requires 380 events. Design A incurs only a marginal increase in the required number of events (to 406). If the putative biomarker is not useful in identifying sensitive patients, design A preserves the power of the overall comparison. On the other hand, if the biomarker is successful in identifying the sensitive subpopulation, design A can provide a substantial increase in power compared with the standard design, which ignores the biomarker (Table 4). For example, if the new treatment produces a 60% reduction in hazard in 25% of the patients, as defined by the biomarker, design A will have 91% power to detect this reduction, whereas the standard design will have 65% power to detect it (see Appendix B for details). If only 10% of the patients are sensitive to the new treatment, as defined by the biomarker, then design A will have 93% power to detect the 60% reduction in hazard, whereas the standard design will have only 35% power to detect this reduction.

When confidence in the ability of the biomarker to identify sensitive patients is reasonably high and preliminary information on the fraction of sensitive patients is available, a more rigorous approach to sample size planning can be adopted (biomarker-adaptive design B; Appendix B). This approach may require a larger sample size than design A because it is considered more likely that the treatment effect will be limited to a subset of patients. However, design B permits the sample size to be reduced compared with a design that ignores the biomarker and evaluates only the diluted overall treatment effect. Table 5 presents the sample size reduction (in terms of the number of events required to detect a 60% reduction in hazard) under a range of fractions of sensitive patients. For example, a broad-eligibility design would need to observe approximately 600 events to have 80% power for detecting a 60% reduction in hazard that is restricted to an undefined subset of 25% of the patients. Design B reduces the required number of events by almost 50% (to 316 events).

Discussion

We have shown that validation of a biomarker for identifying sensitive patients can be prospectively incorporated into randomized phase III trial design without compromising the ability to detect an overall effect. The procedure we have proposed allows drug development to be optimized by combining definitive testing for overall effect with biomarker validation. For the efficiency and validity properties of the procedure to be fully realized, it must be incorporated prospectively in the study protocol.

We used a simplified cutoff model (model [2]) to develop the proposed procedure. This model is based on categorizing the patient population into two groups: sensitive and nonsensitive. Although such categorization results in some loss of information, the resulting test procedure was robust to departures from the cutoff model. The procedure avoids the multiplicity problem (12,13), with selection of an optimal cutoff level by using a permutation resampling approach.

The approach we have presented is directly applicable to predictive indices such as predictive scores based on gene expression microarrays or any other multiple biomarker–based composite indices without a predefined cutoff. The significance-testing component

of our approach is also applicable to settings in which the composite index involves one or more cutoff points for each of several biomarkers. The estimation component of our approach needs to be generalized to the multivariable framework to be used with multiple cut points (see Appendix A for details).

Development and validation of biomarkers to identify patients who are sensitive to new molecularly targeted drugs is a time-consuming process, and thus, reliable biomarkers are often not available at the time definitive phase III studies are designed. If overall treatment effect is not detected in a traditional broad eligibility phase III design, statistically rigorous testing for a subset effect requires a randomized confirmatory study (leading to a considerable increase in cost and duration of the drug development process). For example, Ravaud et al. (14) reported a randomized study of the EGFR/ErbB2 tyrosine kinase inhibitor lapatinib in renal cell cancer patients who express either EGFR or ErbB2. The primary analysis in all patients showed no difference in overall survival (median 46.9 weeks with lapatinib versus 43.1 weeks without it; $P = .29$). Among patients whose tumors overexpressed EGFR, however, median survival was 46.0 weeks in those who received lapatinib versus 37.9 weeks in those who did not receive it ($P = .02$). This result, however, came from a secondary subset analysis that will have to be confirmed in a prospective randomized study. In a setting of this type, our procedure can provide an efficient drug development tool because it combines a formal test for treatment effect in the broad population and a test for a biomarker-defined subset effect in a single clinical trial. In situations where benefit is restricted to a subset of patients, our approach provides a point estimate and confidence interval for the cutoff identifying the sensitive subpopulation. In addition, the estimated probability of benefiting from the new therapy as a function of the biomarker value can be presented to assist in clinical decision making.

If the treatment effect is restricted to a relatively small fraction of the study population (less than 50%), a study that is sized to detect a clinically meaningful overall effect in all randomized patients may have inadequate power. To address this problem, we proposed two biomarker-based trial designs: 1) a simple approach that requires a small increase in sample size for situations in which only limited information on the biomarker is available and 2) a more rigorous approach for situations where preliminary information on the fraction of sensitive patients is available. We showed that the proposed designs provided a substantial gain in efficiency (if the sensitive subpopulation is identified by the biomarker).

In summary, the proposed biomarker-adaptive threshold approach provides a statistically rigorous phase III evaluation of new therapies. The procedure preserves the ability to detect an overall effect if the intervention is beneficial in a broad population. At the same time, the procedure provides a statistically valid test if benefit of the new therapy is limited to a subset of patients as defined by the biomarker. This structured approach is contained within a single randomized clinical trial.

References

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

(2) Simon R, Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. Pharmacogenomics J 2006;6:166–73.

(3) 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.

(4) Dziadziuszko R, Hirsch FR, Varella-Garcia M, Bunn PA Jr. Selecting lung cancer patients for treatment with epidermal growth factor receptor tyrosine kinase inhibitors by immunohistochemistry and fluorescence in situ hybridization—why, when, and how? Clin Cancer Res 2006;12: 4409s–15s.

(5) Cox DR. Regression models and life tables (with discussion). J Royal Stat Soc B 1972;34:187–220.

(6) Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994;86:829–35.

(7) Miller R, Siegmund D. Maximally selected chi-square statistics. Biometrics 1982;38:1011–6.

(8) Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–85.

(9) Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York (NY): Springer; 2004.

(10) Byar DP, Corle DK. Selecting optimal treatment in clinical trials using covariate information. J Chronic Dis 1977;30:445–59.

(11) Andrews DF, Herzberg AM. Data. Chapter 46. New York (NY): Springer; 1985. p. 261–74.

(12) Tukey JW. Some thoughts on clinical-trials, especially problems of multiplicity. Science 1977;198:679–84.

(13) Simon R. Problems of multiplicity in clinical trials. Journal of Statistical Planning and Inference 1994;42:209–221.

(14) Ravaud A, Gardner J, Hawkins R, Von der Maase H, Zantl N, Harper P, et al.; Tykerb Renal Cell Cancer Study Group and GSK CoreT. Efficacy of lapatinib in patients with high tumor EGFR expression: results of a phase III trial in advanced renal cell carcinoma (RCC) 2006 ASCO Annual Meeting Proceedings 2006;24 Abst4502.

(15) Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. J Chronic Dis 1981;34:469–79.

(16) Tsao MS, Sakurada A, Cutz JC, Zhu CQ, Kamel-Reid S, Squire J, et al. Erlotinib in lung cancer—molecular and clinical predictors of outcome. N Engl J Med 2005;353:133–44.

(17) Simon R. Size of phase III cancer clinical trials. Cancer Treat Rep 1985;69:1087–93.

Notes

The authors take full responsibility for the study design, analysis, and interpretation; the writing of the manuscript; and the decision to submit the manuscript for publication.

Manuscript received November 2, 2006; revised April 27, 2007; accepted May 18, 2007.

Appendix A: Statistical methods

For a patient with biomarker value v, a general representation of model [2] is

$$\log h(t) = \log h_0(t) + \mu\tau + \eta I(v > c_0) + \gamma\tau I(v > c_0), \qquad [A1]$$

where μ is the main treatment effect, η is the main biomarker effect, γ is treatment by biomarker interaction, τ is the treatment group indicator ($\tau = 0$ for control arm and $\tau = 1$ for experimental arm), and $I()$ is an indicator function that takes value 0 when $v \le c_0$ and 1 when $v > c_0$. If the main biomarker effect is assumed to be 0, model [2] is obtained.

Note that in our simulations it was assumed that the biomarker values follow a uniform distribution. In many applications, the population distribution of the biomarker is not uniform. To make the distribution uniform, the biomarker values need to be transformed to the percentile scale.

Procedure B is carried out as follows. For values of c in the interval (0, 1), model [2] is fitted on the subset of patients with biomarker values above c. Then the log likelihood ratio statistic $S(c)$ for testing $\gamma = 0$ is calculated. The test statistic is defined as $T = \max((S(0) + R), \max_{0<c<1}\{S(c)\})$. In the second stage of procedure A, $T = \max_{.5<c<1}\{S(c)\}$ is used.

The choice of α_1 (and by implication $\alpha_2 = \alpha - \alpha_1$) for procedure A and the choice of R for procedure B allow one to balance the procedure's ability to detect an overall effect and its ability to detect a subset effect. For procedure A, we chose $\alpha_1 = .04$ to preserve the power of the overall test. At the same time, $\alpha_2 = .01$ was shown to provide a reasonable power against a strong subset effect. Similarly, for procedure B, the choice of R was based on optimizing the ability to detect a subset effect without compromising the overall effect power. We chose $R = 2.2$, which is equal to the difference between the 95th and 80th percentiles from the chi-squared distribution with 1 df.

To adjust for the multiple testing inherent in the construction of statistic T, a resampling-based approach is used (9). First, statistic T is evaluated on the observed data. Then, K permuted datasets are constructed by randomly permuting treatment labels. For each permuted dataset, the corresponding test statistic T^* is calculated. The permutation P value is given by

$$\frac{1+\text{number of permutations where } T < T^*}{1+\text{number of permutations}}.$$

A point estimator $\hat{c}_0$, for the cutoff value c_0 is obtained as

$$\hat{c}_0 = \arg\max_{c_0} l(c_0),$$

where $l(c_0)$ is the partial log likelihood function based on model [A1]:

$$l(c_0) = \max_{\mu\,\eta\gamma} l(\mu, \eta, \gamma, c_0), \quad \text{for } c_0 \in [0,1].$$

An estimate $\hat{F}_{\hat{c}_0}$ of the distribution function $F_{\hat{c}_0}$ of $\hat{c}_0$ is obtained as follows. B random bootstrap samples from the observed data are drawn. For each bootstrap sample, an estimate $\hat{c}_0^*$ is obtained, $\hat{F}_{\hat{c}_0}$ is estimated by the empirical distribution of $\hat{c}_0^*$. For a given biomarker value, function $\hat{F}_{\hat{c}_0}$ gives the estimated probability that the true cutoff level c_0 is below that value. When there is no overall treatment effect, this function is interpreted as the probability that a patient with given biomarker value will benefit from the new therapy. Percentiles of the empirical distribution function provide a confidence interval for c_0. Confidence intervals for $\hat{F}_{\hat{c}_0}$ are obtained by double bootstrap: first B_1 random first-level bootstraps are drawn from the observed data. For each first-level bootstrap sample an estimate of the distribution function $\hat{F}^*_{\hat{c}_0}$ is obtained using B_2 second-level bootstraps. Percentiles of the empirical distribution of $\hat{F}^*_{\hat{c}_0}$ provide a confidence interval for $F_{\hat{c}_0}$. In Figs 1 and 2, $\hat{F}_{\hat{c}_0}$ was obtained using 1000 bootstrap samples; confidence intervals for $\hat{F}_{\hat{c}_0}$ were based on 1000 first-level and 200 second-level bootstrap samples.

The procedures can be modified for a more general setting, where a classifier for sensitive patients is based on multiple cut points and/or multiple biomarkers. Consider a classifier $C_{c_1,\ldots,c_l}(v_1,\ldots,v_k)$ that is defined by l cutoff values $c_1,\ldots,c_l$ and takes values 0 or 1 as a function of k biomarker values $v_1,\ldots,v_k$ (where $v_i \in [0,1]$ for $i = 1,\ldots,k$). Procedures A and B are carried out by fitting model [2] to the subset of patients with the classifier equal to 1 and then maximizing statistic T over all possible cutoff values. The estimation is based on substituting classifier $C_{c_1,\ldots,c_l}(v_1,\ldots,v_k)$ for the single biomarker indicator function $I(v>c_0)$ in model [A1]. The variability of the estimates is represented by a confidence region.

It is widely accepted that, for ethical reasons, randomized clinical trials need to have interim efficacy monitoring. Procedures A and B can be readily adjusted to accommodate interim monitoring. We recommend a conservative monitoring procedure that preserves most of the type I error for the final analysis. For procedure A, interim monitoring can be performed with respect to the overall test controlling the significance level at α_1. The subset test is performed at significance level α_2. The overall significance level of the procedure $\alpha = \alpha_1 + \alpha_2$ is preserved. For procedure B, we also recommend using the overall test for interim monitoring. If the study is not stopped early for strong overall effect, procedure B is applied at the final analysis using the balance of the type I error.

Appendix B: Sample Size for Treatment Effect Restricted to a Subpopulation

In a standard broad-eligibility design with time-to-event outcomes, the sample size calculation is usually based on the number of observed events required to detect the target treatment effect in all randomly assigned patients (at significance level α) with power $1 - \beta$ (where β denotes the false-negative error rate). The target treatment effect is usually expressed in terms of the ratio of the hazard in the experimental arm over that in the control arm $\Delta = h_E/h_C$. The total required number of events, D, is approximately

$$D = 4\left(\frac{C_{1-\alpha} + C_{1-\beta}}{\log \Delta}\right)^2, \qquad [B1]$$

where $C_{1-\alpha}$ and $C_{1-\beta}$ denote percentiles of the standard normal distribution (15).

If treatment effect Δ is limited to a fraction π of the population, it can be shown (1) that the number of events D_S required for the overall test is approximately

$$D_S \approx 4\left(\frac{C_{1-\alpha} + C_{1-\beta}}{\pi \log(\Delta)}\right)^2 = D/\pi^2. \qquad [B2]$$

That is, the number of events required for the standard design when treatment effect is limited to a fraction π of the population is the number of events that would be required if an effect of the same magnitude applied to the entire study population divided by π^2.

Description of Biomarker-Adaptive Design B. If there is reasonable confidence that treatment effect may be limited to a biomarker-defined subset of patients and there is preliminary information on the fraction, π, of sensitive patients, the sample size for clinical trials using procedure B is derived in two steps. First, the relative power (relative efficiency) of procedure B relative to the broad eligibility design is estimated from Table 1 by the ratio of column 4 over column 6 for the row corresponding to the expected proportion of sensitive patients π and the design hazard ratio Δ. This relative efficiency is used to calculate the power of the overall test corresponding to the desired power of procedure B. In the second step, formula B2 is used to estimate the number of events needed for the overall test to have this power. For example, for $\pi = 25\%$ and $\Delta = 0.4$ (a 60% reduction in hazard), the relative efficiency is 0.429/0.641 = 0.67 (from simulation 4 line 2). Therefore, when procedure B has power of 80% to detect $\Delta = 0.4$, the corresponding overall effect test has power 80% ×.67 = 53%. From formula B2, approximately 316 events are needed to have 53% power with $\Delta = 0.4$ and $\pi = 25\%$ ($\alpha = 0.05$, two-sided). Thus, a study using procedure B will need 316 events to have 80% power to detect the treatment effect ($\Delta = 0.4$) limited to 25% of the population. By contrast, a broad-eligibility study requires 600 events for the same power.

In general, it is reasonable to expect a larger treatment effect from a targeted drug in the sensitive subpopulation than the effect sizes used in designing broad-eligibility trials. For example, Tsao et al. (16) reported an HR of 0.44 for the tyrosine kinase inhibitor erlotinib versus placebo in the subset of non–small-cell lung cancer patients with high EGFR gene copy number. This hazard ratio corresponds to a substantially higher treatment effect than the hazard ratios that are typically used in design of broad-eligibility trials in cancer (e.g., an HR of 0.75). Thus, it is reasonable to use the target subset HR of 0.4 in the sample size calculation for the procedure B–based design. The design is illustrated in Table 5. It presents the number of events required for 80% power to detect a HR of 0.4 for (1) a standard broad-eligibility design, and (2) a trial using procedure B, for fractions of sensitive patients $\pi = 75\%$, 50%, 25%, and 10%.

Table 4 (for biomarker-adaptive design A) was obtained through the relative efficiency calculation using Table 1 as described above.

Once the required number of events is determined, the total number of patients needed is calculated as the required number of events divided by the expected average event rate. Where the expected average event rate is a function of the average hazard rate in two arms $h = (h_E + h_C)/2$, the duration of follow-up period F and the duration of accrual period A: $r = (1 - \frac{e^{-Fh}}{Ah}(1 - e^{-Ah}))$. For more details, see Simon (17).