VOLUME 23 · NUMBER 9 · MARCH 20 2005

JOURNAL OF CLINICAL ONCOLOGY

REVIEW ARTICLE

Clinical Trial Designs for Predictive Marker Validation in Cancer Treatment Trials

Daniel J. Sargent, Barbara A. Conley, Carmen Allegra, and Laurence Collette

From the Mayo Clinic, Rochester, MN; Medical Oncology Clinical Research Unit, Center for Cancer Research, National Cancer Institute, Bethesda, MD; Network for Medical Communication and Research, Atlanta, GA; and Data Center, European Organisation for Research and Treatment of Cancer, Brussels, Belgium.

Submitted January 16, 2004; accepted September 27, 2004.

Supported by and prepared under the auspices of the Joint National Cancer Institute/European Organization for Research and Treatment of Cancer Working Group on Clinical Trials Methodology for Tumor Marker Studies.

Authors' disclosures of potential conflicts of interest are found at the end of this article.

Address reprint requests to Daniel J. Sargent, PhD, Mayo Clinic, Kahler 1A, 200 First St, SW, Rochester, MN 55905; e-mail: sargent.daniel@mayo.edu.

0732-183X/05/2309-2020/$20.00

DOI: 10.1200/JCO.2005.01.112

ABSTRACT

Current staging and risk-stratification methods in oncology, while helpful, fail to adequately predict malignancy aggressiveness and/or response to specific treatment. Increased knowledge of cancer biology is generating promising marker candidates for more accurate diagnosis, prognosis assessment, and therapeutic targeting. To apply these exciting results to maximize patient benefit, a disciplined application of well-designed clinical trials for assessing the utility of markers should be used. In this article, we first review the major issues to consider when designing a clinical trial assessing the usefulness of a predictive marker. We then present two classes of clinical trial designs: the Marker by Treatment Interaction Design and the Marker-Based Strategy Design. In the first design, we assume that the marker splits the population into groups in which the efficacy of a particular treatment will differ. This design can be viewed as a classical randomized clinical trial with upfront stratification for the marker. In the second design, after the marker status is known, each patient is randomly assigned either to have therapy determined by their marker status or to receive therapy independent of marker status. The predictive value of the marker is assessed by comparing the outcome of all patients in the marker-based arm to that of all of the patients in the non–marker-based arm. We present detailed sample size calculations for a specific clinical scenario. We discuss the advantages and disadvantages of the two trial designs and their appropriateness to specific clinical situations to assist investigators seeking to design rigorous, marker-based clinical trials.

J Clin Oncol 23:2020-2027.

INTRODUCTION

Current staging and risk-stratification methods for malignant disease incompletely predict prognosis and/or treatment efficacy. As new therapeutic options emerge, it is desirable to use our increasing knowledge of tumor molecular biology to optimize and individualize therapy. Reports from exploratory studies regularly suggest potentially useful candidate markers for this purpose. However, few markers are currently developed to the point of allowing reliable use in clinical practice. The lack of a disciplined approach will slow the introduction of markers into clinical use, or alternatively, markers may be introduced without sufficient scientific evidence of benefit. Clinical trial designs for evaluating the usefulness of molecular traits or markers within the context of treatment trials in cancer patients is the subject of this article.

DEFINITIONS AND PRELIMINARY DATA REQUIRED

We refer to a marker as a property of the tumor associated with a clinical outcome. It may be a single trait, or a grouping (signature) of traits that separates different populations with respect to an outcome of interest. Prognostic markers classically identify patients with differing risks of a specific outcome, such as progression or death.[1,2] Because the clinical scenarios where no effective treatment options exist have become rare, we take a slightly expanded definition, defining a prognostic marker as one that informs about outcome in the absence of systemic therapy or portends an outcome different from that of patients without the marker, despite empiric (not targeted to the marker) systemic therapy. For example, under this definition, a marker would be prognostic if, in the

absence of adjuvant therapy, patients with tumors expressing a specified marker have a poorer survival compared with patients without the marker. The uniform introduction of adjuvant therapy may improve the survival of the whole group, but if survival of patients expressing the marker remains poorer than that of marker-negative patients, the marker remains prognostic under this definition (Tables 1 through 5). We choose this definition because the marker is associated with a differential outcome regardless of the therapy given, even if a choice of therapy is available. A prognostic marker can distinguish populations into groups where different treatment options are appropriate (possibly including no treatment), but it cannot guide the choice of a particular therapy. The preliminary knowledge necessary to propose a validation trial of a prognostic marker is the subject of considerable previous work.[1-5]

A predictive marker is a marker that predicts the differential efficacy (benefit) of a particular therapy based on marker status (eg, only patients expressing the marker will respond to the specific treatment or will respond to a greater degree than those without the marker). A predictive marker could, therefore, guide the choice of therapy. Such predictive markers could be relevant to the choice of therapy in one of several ways. For example, if a marker (eg, a receptor such as HER2/*neu* [for trastuzumab] or c-kit [for imatinib mesylate]), is linked to the development or course of the particular malignancy, then the presence or number of receptors is a promising marker when a potential therapeutic agent targeted to the receptor is developed. Examples of the impact of a predictive marker on treatment decisions are listed in Tables 1 through 5.

Before development of a prognostic or a predictive marker for clinical use, the relationship between marker expression and outcome should be explored by retrospective study of the marker in available tissues of patients with known outcome who have been treated similarly. Complete data on potentially confounding factors allow a more convincing preliminary assessment. A marker that is independently associated with outcome after adjusting for such factors is considered promising for clinical utility. For the purpose of designing a validation trial, information regarding the prognostic properties of the putative marker relative to existing (standard) treatment, as well as the predictive effect of the marker relative to the new (targeted) treatment, is necessary to propose specific, testable hypotheses. Once the marker meets the criterion of promising, additional data must be gathered before initiating confirmatory studies to test its clinical utility. These data include the specificity of the marker to the cancer of interest (as opposed to normal tissues, other disease states, or other cancers), an estimate of the marker prevalence in the target population, confidence in the method of measurement, including definition of any cut points, and demonstration that the measurement can be reliably performed on the specimens that will be available.

Assuming that a marker (or a group of markers, such as a molecular signature) has met the development milestones described, an evaluation of clinical utility requires the selection of the appropriate patient population and the choice of the most appropriate end point. Ideally, the population studied should be one in which knowledge of the marker would have substantial clinical relevance and where the feasibility of obtaining appropriate specimens is established. For example, it may be quite feasible to obtain tumor tissue (from initial resection) from patients in a trial evaluating the predictive ability of a marker for a specific adjuvant therapy. However, a similar trial in patients with metastatic disease may require a biopsy of metastatic tumor unless it has been established that the characteristics of the metastases are substantially similar to the (usually more available) resected primary tumor. This requirement may reduce the number of patients willing to enter the trial, increase patient risk, and perhaps introduce bias by skewing the population to those needing or willing to have a biopsy.

The choice of primary end point is also critical. In evaluating predictive markers of therapeutic efficacy in the adjuvant setting, the primary end point will usually be

Table 1. Prognostic Marker: No Difference Between A and B in the Population

	Median Survival (months)		
Marker Status	No Treatment	Treatment A	Treatment B
High	3	6	6
Low	6	12	12

NOTE. High marker levels are prognostic of worse outcome irrespective of treatment (hazard ratio of 0.5 comparing the two marker levels).

Table 2. Prognostic Marker: Treatment B Is More Effective Than Treatment A (hazard ratio = 1.5) in the Population and the Marker Is Prognostic in All Groups (hazard ratio = 0.5)

	Median Survival (months)		
Marker Status	No Treatment	Treatment A	Treatment B
High	3	6	9
Low	6	12	18

Table 3. Predictive Marker: The Marker Is Not Prognostic, But Low Marker Levels Are Predictive of Better Outcome From Treatment B (hazard ratio for treatment = 0.5)

	Median Survival (months)		
Marker Status	No Treatment	Treatment A	Treatment B
High	6	9	9
Low	6	9	18

Table 4. Predictive Marker: Marker Level Is Prognostic Irrespective of Treatment (hazard ratio = 0.5) and Predictive for Better Outcome From Treatment B (hazard ratio = 0.5) in Patients With Low Marker Levels

	Median Survival (months)		
Marker Status	No Treatment	Treatment A	Treatment B
High	3	4.5	4.5
Low	6	9	18

overall, disease-free, or recurrence-free survival. Possible primary end points for metastatic disease trials include response rate, time to progression, survival, or risks of toxicity. Time to progression may be a good surrogate end point for efficacy in many diseases, provided careful attention is paid to uniform measurement of this parameter. Survival is the gold standard, but in many cases it may be affected by second- and third-line treatments.[6] Clearly, the trial sample size and duration needed to achieve the desired goal should be kept as small as possible. A trial of long duration risks the possibility that results are so delayed that relevance is compromised; the standard of treatment may change, thus forcing a decision to either stop the trial early or to change the treatment. A series of trials may be needed and compromises necessary between what would be ideal and what is practicably achievable.

CLINICAL TRIAL DESIGNS FOR ASSESSMENT OF THE CLINICAL UTILITY OF A PREDICTIVE MARKER

Clinical trial design methodologies to test the clinical usefulness of putative prognostic or predictive factors could best be described as maturing. Here we review four designs for predictive marker studies.

Indirect Assessment (Fig 1)

1. Marker by Treatment Interaction Design, separate tests. In this design, we assume the marker splits the population into two groups. Patients in each marker group are randomly assigned to two different treatments, and the testing plan determines whether one treatment is superior to the other separately within each marker group.

2. Marker by Treatment Interaction Design, test of interaction. In this design, we again separate the population based on marker status and seek to determine whether the treatment effect seen in one group differs significantly from the treatment effect seen in the other group by a formal statistical test for interaction between marker status and treatment assignment.

In essence, these designs undertake two independent clinical trials of treatment A versus treatment B, one in each of two patient groups defined by marker status.

Table 5. Predictive Marker: Marker Level Is Prognostic for Outcome Irrespective of Treatment (hazard ratio = 0.5) and Predictive for Response to Treatment B (greater benefit to high marker patients)

	Median Survival (months)		
Marker Status	No Treatment	Treatment A	Treatment B
High	3	4.5	9
Low	6	9	9

Direct Assessment (Fig 2)

3. Marker-Based Strategy Design (Fig 2A). In this design, after the marker status is known, each patient is randomly assigned to either have his/her therapy determined by their marker status or to receive therapy independent of marker status.[7] The determination of the marker status before randomization ensures nearly 100% availability of the marker in the randomized sample. Figure 2A illustrates the simplest application of this design, in which patients are randomly assigned to either a marker-based or a non–marker-based arm. All patients in the non–marker-based arm receive the same treatment (ie, standard treatment A), whereas patients in the marker-based arm receive treatment A if their marker status is negative and an experimental treatment B (likely a marker-based treatment) if their marker status is positive. The predictive value of the marker is assessed by comparing the outcome of all of the patients in the marker-based arm to that of all of the patients in the non–marker-based arm.

4. Modified Marker-Based Strategy Design (Fig 2B). The Marker-Based Strategy Design described in Fig 2A does not examine the effect of the marker-based treatment in patients with negative marker status, as none of those patients receive that treatment. If the marker-based treatment were superior in all patients, regardless of their marker status, this could not be determined with the application of the design shown in Fig 2A. In a modified version of the Marker-Based Strategy Design, patients in the non–marker-based arm undergo a second randomization to receive one of the same two treatments being used in the marker-based arm. This modification allows clarification of whether any finding regarding the efficacy of the marker-directed approach to therapy is due to a true effect of marker status or to an improved regimen regardless of marker status. This design also may allow a retrospective assessment of an alternative classification for the marker.[7] Ideally, the randomization plan should take account of marker prevalence in the population, as well as of marker status.

Example

In this section we consider practical issues for a trial testing the value of a predictive marker and provide specific sample size estimates for the application of the design to a specific clinical scenario, that of assessing the utility of thymidylate synthase (TS) expression as a predictor for the

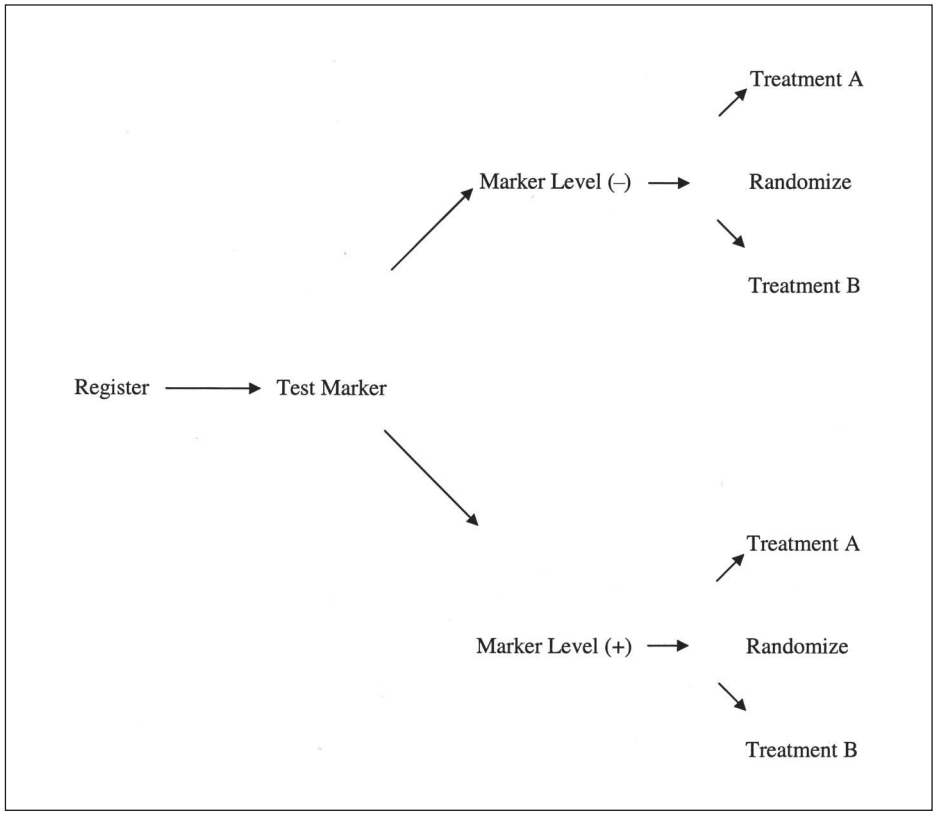

Fig 1. Marker by Treatment Interaction Design to test a predictive factor question; same treatments in both prognostic groups.

efficacy of fluorouracil (FU) -based treatment for colon cancer. FU is metabolized to 5-fluorodeoxyuridine monophosphate, which binds competitively to TS, preventing synthesis of thymidine and of DNA.[8] Exploratory studies suggest that FU is more efficacious for patients whose tumors exhibit low TS levels.[8-14] High TS level may also have an adverse prognostic effect: patients with high TS levels have been shown to have a poorer outcome in the absence of treatment.[9,10] Here we consider the design of a trial to test the hypothesis that allocation of patients to treatment based on tumor TS level improves patient outcome. Specifically, we consider a trial where the two arms are irinotecan plus oxaliplatin (IO),[15] and irinotecan plus FU with leucovorin (IFL).[16] Under our hypothesis, one would expect the FU-containing treatment (defined as treatment B) to be most effective in patients with low TS levels, whereas the efficacy of the non–FU-containing treatment (treatment A) would be unaltered by TS level, except that patients with high expression of TS may have a slightly worse outcome as a result of the small prognostic effect of TS.

For this example, we assume that 50% of patients will have high TS levels and that survival follows an exponential model. In the unselected population with metastatic colon cancer, both treatment regimens result in median survivals of approximately 15 months.[15,16] However, if patients with low TS levels respond preferentially to FU, one might expect a median survival for low TS patients treated with IFL (containing FU) of 20 months, whereas patients with high TS levels receiving IFL may have a median survival of only 12 months. The efficacy of treatment with IO (not containing FU) should be independent of TS level, and thus we might assume that when treated with IO, low TS patients have a median survival of 16 months and high TS patients have a median survival of 14 months, with the difference owing to the prognostic effect of TS. With these assumptions, in the high TS patients, IFL treatment is inferior to IO (with a hazard ratio of 0.86 [12 *v* 14 months median survival] in favor of IO), whereas in the low TS patients, the outcome is superior with IFL compared with IO (with a hazard ratio of 1.25 in favor of IFL [20 *v* 16 months median survival]). The resulting hazard ratio for the interaction between treatment and marker level is 0.69 (0.86 *v* 1.25). Considering the Modified Marker-Based Strategy Design in this setting, the size of the strategy effect (ie, using a marker-based rather than a non–marker-based strategy for choosing therapy) is limited: the overall median survival in the marker-based arm is 16.5 months, compared with 15 months in the non–marker-based treatment group, resulting in a hazard ratio comparing the two arms in the primary randomized comparison (marker-based *v* non–marker-based arm) of 0.91 (15 months *v* 16.5 months). This relatively small effect is related to the rather limited treatment

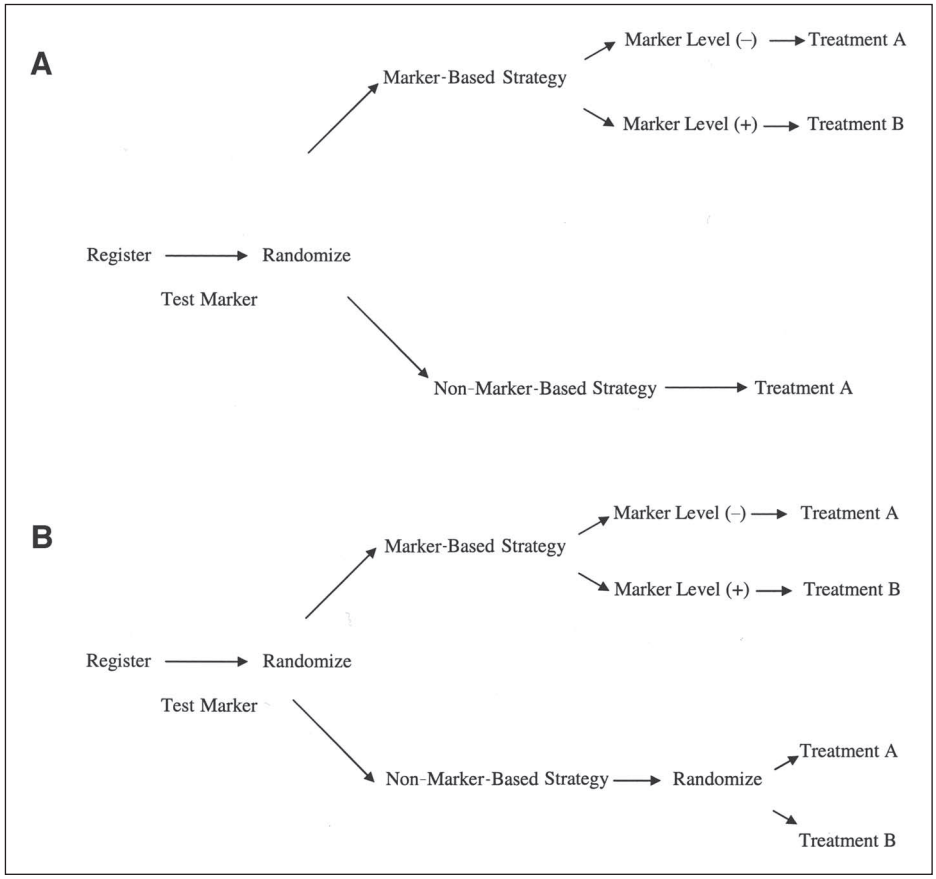

Fig 2. (A) Marker-Based Strategy Design to test predictive factor question; no randomization in non–marker-based arm. (B) Marker-Based Strategy Design to test predictive factor question; randomization in both arms.

effect in reach of the two marker groups (hazard ratio = 0.86 and 1.23, respectively), and the fact that in the non–marker-directed therapy arm, 50% of the patients receive the optimal therapy by chance alone.

The required sample sizes for 90% power to detect these hypothesized differences, using the Marker by Treatment Interaction (Fig 1) design with analysis by separate tests or by test of interaction, as well as the Modified Marker Strategy design (Fig 2B) for this example, are shown in Table 6. From this Table, we see that for the Marker by Treatment Design, using separate tests (patients separated by marker status and then randomly assigned to either IO or IFL), we need to observe 1,705 events in the high TS arm and 844 events in the low TS arm (2,549 events in total) to have 90% power to detect a meaningful effect. To observe this number of events, we would need to enroll 2,756

Table 6. Sample Sizes Required for TS Example, 50% Prevalence, 90% Power, Two-Sided $\alpha = .05$

Comparison	Total No. of Events Required in Trial	No. of Patients Required/Arm (assuming a mean 18 months of follow-up)
Marker by Treatment Interaction Design		
Comparison of IO and IFL within high-TS patients	1,705	1,378
Comparison of IO and IFL within low-TS patients	844	845
Total		
Separate tests	2,549	2,223
Test of interaction	1,220	1,048
Modified Marker Strategy Design	4,629	4,215

NOTE. The assumptions for the median overall survival (in months) are: thymidylate synthase (TS; low (L)/IFL, 20 months; TS (L)/IO, 16 months; TS (high [H])/IFL, 12 months; TS (H)/IO, 14 months. Hazard ratio (IFL/IO - TS [H]): 0.86; hazard ratio (IFL/IO- TS [L]): 1.25; hazard ratio (interaction): 0.69; hazard ratio (marker-based strategy arm/non–marker-based strategy arm): 0.91.
Abbreviations: IO, irinotecan plus oxaliplatin; IFL, irinotecan plus fluorouracil with leucovorin.

patients in the high marker group (1,378 assigned to IFL and 1,378 assigned to IO) and 1,690 patients in the low marker group (thus we note that accrual could be terminated early in the low TS group, assuming 50% prevalence). If the analyses were by test of interaction, 1,220 events would need to be observed, requiring that 1,048 patients per arm (1,048 to IFL, 1,048 to IO, regardless of marker status) would need to be enrolled. For the Modified Marker Strategy Design in which patients are randomly assigned to be treated by marker status or not, we would need to observe 4,629 events and to enroll 4,215 patients in each arm (4,215 patients where treatment was assigned based on marker, the same number where treatment is not marker-dependent).

The sample size necessary to answer the relevant questions using the Treatment by Marker Interaction Design (Fig 1) is smaller than that necessary to answer the main question using the Modified Marker Based Strategy Design 4. In this example, the hazard ratio for the Modified Marker Strategy Design (Fig 2B) is very close to 1.00; therefore, this design necessitates a much larger sample size than the Treatment by Marker Interaction Design when used to test for a formal interaction (design 2) or even the indirect design if used to conduct separate tests (design 1). In this example and others we have investigated, the detection of the interaction effect requires a sample size smaller than the total sample size needed to provide adequate power to answer the treatment question within each marker group individually. The reason for this fact is that the interaction effects under investigation are at least as large as the smallest target treatment effect in the subset.

In general, as demonstrated by this example, the sample sizes for assessing the clinical utility of the putative predictive marker in this example are quite large. This results from at least three factors. First, the assumed marker effects were modest. However, this modest hazard ratio is similar to the hazard ratios used to design phase III treatment trials and may indeed be clinically relevant, in that changes in practice are often based on such differences in hazard ratios. Second, in this example, the putative marker was predictive for only one of the regimens, in this case, IFL. The predictive importance of the marker on the other regimen, IO, was assumed to be null. If a marker was hypothesized to be predictive for multiple complimentary regimens, the sample sizes could be greatly reduced (see Appendix for an example of such a scenario). The final reason for the large sample sizes is unavoidable: investigation of predictive effects for a marker is, by definition, a prospective subset analysis: in other words, does the treatment effect differ in subgroups defined by a marker level? Therefore, a larger sample size is necessary.

DISCUSSION

The choice of design for any particular trial depends on the nature of the conclusions that wish to be drawn and the strength of evidence desired at the trial's conclusion. For the Marker by Treatment Interaction Design using separate tests (design 1), the trial's sample size (and thus its power) is based on testing the treatment effect separately in the two marker groups. With such testing, the result may be that there is a statistically significant benefit to one treatment in neither, one, or both marker groups. Such an approach may or may not provide convincing evidence of a marker's utility. Clearly, if treatment A is superior in one marker group and treatment B is superior in the other, the marker has clinical utility. However, if these same trends are observed, but are nonsignificant, the clinical utility of the marker remains unclear. The Marker by Treatment Interaction Design using a test of interaction (design 2) addresses this issue by basing the sample size on having adequate power to test for a differential treatment effect in the two marker groups. In this approach, we can test with adequate power that any difference in treatment benefit observed in the two marker groups is itself statistically significant. This approach has the advantage of using all randomly assigned patients in a single test, thus maximizing efficiency. However, it does not provide power for testing the treatment effect separately in the two marker subgroups. We note briefly that this design also allows for an evaluation of the prognostic value of the marker in patients treated with regimens A and B by comparing the outcomes of patients treated with the same regimen between the two marker groups.

On the basis of the example used in this article, one would most likely initially choose one of the Marker by Treatment Interaction Designs in the instances of a single marker dictating a choice between two treatments to assess the clinical utility of a putative predictive marker, despite the drawbacks of the indirect assessment noted above. This design could be powered to either detect a given difference in treatment outcome within patients in each of two marker groups (separate tests) or to detect an interaction between the marker value and the treatment efficacy (test of interaction) to demonstrate the usefulness of a marker for choosing a particular therapy. Although this design is efficient, it is a less direct test than that presented in the Marker-Based Strategy Designs. This trade-off needs to be considered in the design of any marker-based trial and will be based in part on the practicality of obtaining marker information.

The currently accruing European Organization for Research and Treatment of Cancer Trial 10994 is using the Marker by Treatment Interaction Design (design 1). In this trial, patients with breast cancer are classified by whether the tumor has a normal or mutated p53 protein. Patients are randomly assigned to either an anthracycline-based regimen only or to a taxane plus anthracycline regimen, with the same randomization for patients with normal and mutated p53 tumors. Preliminary data suggest that patients with mutated p53 respond less well to anthracyclines compared with patients with normal p53. The hypothesis being

tested is whether the addition of a taxane (for which response is presumed to be independent of p53 mutation status) will improve the outcome for patients with mutated p53.[17,18] If the results of the trial confirm this hypothesis, then it may be reasonable to choose a taxane-containing regimen in patients with tumors that have mutated p53, whereas in patients with normal p53, the addition of a taxane may not be necessary.

In other arenas, the Marker-Based Strategy Design has significant merit. For example, if a treatment decision is to be based on a panel of markers, if there are more than two treatments to which patients can be assigned, or if other outcomes in addition to efficacy should be considered, implementation of the Marker by Treatment Interaction Design is problematic or even impossible. However, the Marker-Based Strategy Design (Fig 2A) may be applied to randomize between marker-directed versus non–marker-directed treatment. This design variation is being used in a current clinical trial in platinum-resistant recurrent ovarian carcinoma. In that trial, patients are randomly assigned to have therapy determined among 11 possible regimens by either a marker assay (an adenosine triphosphate–based chemosensitivity assay) or physician choice.[19] As molecular evaluation becomes more sophisticated, outcomes in addition to treatment efficacy (response or survival) may also be able to be evaluated in this way. Such possibilities include the evaluation of impact of genetic polymorphisms on severe toxicity or the minimization of toxicity when efficacy is equivalent. We expect that as targeted therapies become increasingly available, and as predictive approaches become more sophisticated (including molecular profiling), the marker-based strategy design will become increasingly attractive. We also note that both versions of the Marker-Based Strategy Design (Figs 2A and 2B) also permit an assessment of the prognostic value of the marker in question by comparing outcome in patients by marker level who receive standard treatment.

In special cases, to reduce a trial's sample size, one could consider a partial strategy of assessing the clinical utility of a marker. For example, for a marker trial embedded in a treatment trial (phase III comparison of two treatments), the end point for the marker trial could be response rate, whereas for the treatment trial, the end point would be overall survival. This tactic assumes that any improvement in survival associated with the use of a given therapeutic strategy would occur in the responding rather than the nonresponding patients. However, such options should only be used for cases where it would be impractical to do a more rigorous clinical evaluation of the marker.

One might consider the possibility of retroactively applying the Marker by Treatment Interaction Design (Fig 1) at the conclusion of a usual randomized phase III cancer treatment clinical trial; that is, retrospectively assess tumor specimens from a completed trial, classify patients into groups based on their marker status, and compare the two treatments separately in the two marker groups. This approach may be useful for demonstrating the clinical utility of a marker, but proper design should be used. Because it is unlikely that tumor specimens will be available on 100% of the enrolled patients, the study may be suboptimal because of the possibility for bias in the samples that are available and in the lower statistical power to discern an effect. However, if the marker-based analyses were planned prospectively and tissue were available on all or most patients, adequate statistical power may exist to compare the two treatments separately in the two marker groups.

One strategy to answer a marker-based question more feasibly may be to consider a trial with two patient cohorts: a primary cohort to answer the treatment question and a second cohort to answer the marker-based question. The extensive data needed for the treatment trial (including dose intensity, toxicity, and so on) may not be essential to address the marker-based question, which may only require data on vital status (in addition to the marker information). The addition of a second cohort of patients to a typical treatment trial, with the collection of a significantly reduced data set, may provide a cost-effective approach to allow marker-based questions to be addressed.

In a setting where there is only one marker and one marker-based (presumably experimental) treatment available, these marker-based designs may not be necessary. It may be more efficient to determine those patients who have the marker, randomly assign them to either standard treatment or the marker-based treatment, and determine the efficacy of the new treatment for marker-positive patients. In this case, the marker negative patients will not participate in the evaluation, but will be treated separately. However, such an approach provides no information on treatment efficiency in marker-negative patients. Given the lack of relationship between expression levels of a target and the efficacy of some targeted agents,[20,21] such an approach should be applied cautiously.

In this article, we considered trial designs for evaluation of a single marker. As new technologies emerge, multiple markers, or a comprehensive set of markers, defining a distinct tumor behavior, may need to be evaluated. If a combination of markers can be treated as a unit, or classifier, then clinical trial designs similar to those described above could be considered. However, in cases where markers are assessed independently, the study design would need to be rigorous and sufficiently robust to accommodate multiple comparisons, and the required sample size would likely be larger. The efficiency of these clinical trial designs used to evaluate markers should continue to be evaluated. Novel trial designs should also be tested so that the promise of tumor molecular biology can be translated efficiently to the clinic.

Acknowledgment

We thank Lisa McShane, PhD, Richard Kaplan, MD, and two referees for careful review and helpful suggestions that greatly improved this manuscript.

Appendix

Here we briefly consider a circumstance when a marker is proposed to be predictive for two treatment options in a complimentary manner. Specifically, we consider the case where patients with a marker level considered to be high have a more favorable response to a treatment A, whereas patients with a marker level considered to be low have a favorable response to treatment B. This would be the case in the example presented (TS), if there were a treatment hypothesized to have greater efficacy in patients with high TS levels (recalling that IFL is hypothesized to work best in patients with low TS levels). Here we assume such a treatment has differentiated efficacy of the same magnitude as IFL; that is, the new treatment (which we define as treatment X) provides a median survival of 12 months in low TS patients and 20 months in high TS patients. These assumptions (with the continued assumption of 50% prevalence of high TS *v* low TS), result in the following hazard ratios: HR (IFL/X-high TS): 1.66, HR (IFL/X-low TS): 0.6, HR (interaction): 2.78, HR (marker-based strategy/non–marker-based strategy): 1.25.

The divergence of these various hazard ratios from unity substantially decreases the required sample sizes for the validation clinical trial compared with those presented in Table 2. For example, the required number of events for the Marker by Treatment Interaction Design (separate tests) is reduced eight-fold, from 2,549 to 322; the number of events for the Marker by Treatment Interaction Design (test of interaction) is reduced nine-fold, from 1,220 to 132, and the number of events required for the Modified Marker Based Strategy Design is reduced five-fold, from 4,629 to 844.

Authors' Disclosures of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

REFERENCES

1. Simon R, Altman DG: Statistical aspects of prognostic factor studies in oncology. Br J Cancer 69:979-985, 1994

2. Hayes DF, Trock B, Harris AL: Assessing the clinical impact of prognostic factors: When is "statistically significant" clinically useful? Breast Cancer Res Treat 52:305-319, 1998

3. 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 88:1456-1466, 1996

4. Hammond MEH, Taube SE: Issues and barriers to development of clinically useful tumor markers: A development pathway proposal. Semin Oncol 29:213-221, 2002

5. McGuire WL: Breast cancer prognostic factors: Evaluation guidelines. J Natl Cancer Inst 83:154-155, 1991

6. Di Leo A, Bleiberg H, Buyse M: Overall survival is not a realistic end point for clinical trials of new drugs in advanced solid tumors: A critical assessment based on recently reported phase III trials in colorectal and breast cancer. J Clin Oncol 21:2045-2047, 2003

7. Sargent D, Allegra C: Issues in clinical trial design for tumor marker studies. Semin Oncol 29:222-230, 2002

8. Longley DB, Harkin DP, Johnston PG: 5-fluorouracil: Mechanisms of action and clinical strategies. Nat Rev Cancer 3:330-338, 2003

9. Johnston PG, Fisher ER, Rockette HE, et al: The role of thymidylate synthase expression in prognosis and outcome to adjuvant chemotherapy in patients with rectal cancer. J Clin Oncol 12:2640-2647, 1994

10. Cascinu S, Graziano F, Valentini M, et al: Vascular endothelial growth factor expression, S-phase fraction and thymidylate synthase quantitation in node-positive colon cancer: Relationships with tumor recurrence and resistance to adjuvant therapy. Ann Oncol 12:239-244, 2001

11. Pullarkat ST, Stoehlmacher J, Ghaderi V, et al: Thymidylate gene polymorphism determines response and toxicity of 5-FU chemotherapy. Pharmacogenomics J 1:65-70, 2001

12. Etienne M-C, Chazal M, Laurent-Puig P, et al: Prognostic value of tumoral thymidylate synthase and p53 in metastatic colorectal cancer patients receiving fluorouracil-based chemotherapy: Phenotypic and genotypic analyses. J Clin Oncol 20:2832-2843, 2002

13. Leichman CG, Lenz HJ, Leichman L, et al: Quantitation of intratumoral thymidylate synthase expression predicts for disseminated colorectal cancer response and resistance to protracted-infusion fluorouracil and weekly leucovorin. J Clin Oncol 15:3223-3229, 1997

14. Aschele C, Debernardis D, Casazza S, et al: Immunohistochemical quantitation of thymidylate synthase expression in colorectal cancer metastases predicts for clinical outcome to fluorouracil-based chemotherapy. J Clin Oncol 17:1760-1770, 1999

15. Wasserman E, Cuvier C, Lokiec F, et al: Combination of oxaliplatin plus irinotecan in patients with gastrointestinal tumors: Results of two independent phase I studies with pharmacokinetics. J Clin Oncol 17:1751-1759, 1999

16. Saltz LB, Cox JV, Blanke C, et al: Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. N Engl J Med 343:905-914, 2000

17. Borresen-Dale A-L: TP53 and Breast Cancer. Hum Mutat 21:292-300, 2003

18. Kandioler-Eckersberger D, Ludwig C, Rudas M, et al: TP53 mutation and p53 overexpression for prediction of response to neoadjuvant treatment in breast cancer patients. Clin Cancer Res 6:50-56, 2000

19. Kurbacher CM, Untch M, Cree IA: Protocol 97PRT/1: A randomised trial of chemotherapy directed by a tumour chemosensitivity assay versus physician's choice in patients with recurrent platinum-resistant ovarian adenocarcinoma (accepted protocol). Lancet http://www.thelancet.com/info/info.isa?n1=authorinfo&n2=Protocol+review&uid=1185

20. Cunningham D, Humblet Y, Siena S, et al: Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. N Engl J Med 351:337-345, 2004

21. Kris MG, Natale RB, Herbst RS, et al: Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer: A randomized trial. JAMA 290:2149-2158, 2003