Published in final edited form as:
Per Med. 2010 January 1; 7(1): 33–47. doi:10.2217/pme.09.49.

Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology

Richard Simon
National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA, Tel.: +1 301 496 0975, Fax: +1 301 402 0560

Richard Simon: rsimon@nih.gov

Abstract

Physicians need improved tools for selecting treatments for individual patients. Many diagnostic entities hat were traditionally viewed as individual diseases are heterogeneous in their molecular pathogenesis and treatment responsiveness. This results in the treatment of many patients with ineffective drugs, incursion of substantial medical costs for the treatment of patients who do not benefit and the conducting of large clinical trials to identify small, average treatment benefits for heterogeneous groups of patients. In oncology, new genomic technologies provide powerful tools for the selection of patients who require systemic treatment and are most (or least) likely to benefit from a molecularly targeted therapeutic. In the large amount of literature on biomarkers, there is considerable uncertainty and confusion regarding the specifics involved in the development and evaluation of prognostic and predictive biomarker diagnostics. There is a lack of appreciation that the development of drugs with companion diagnostics increases the complexity of clinical development. Adapting to the fundamental importance of tumor heterogeneity and achieving the benefits of personalized oncology for patients and healthcare costs will require paradigm changes for clinical and statistical investigators in academia, industry and regulatory agencies. In this review, I attempt to address some of these issues and provide guidance on the design of clinical trials for evaluating the clinical utility and robustness of prognostic and predictive biomarkers.

Keywords

adaptive design; biomarker; clinical trial design; predictive; prognostic; validation

The dominant themes in oncology therapeutics today are the molecular heterogeneity of tumors of a common primary site, the development of drugs that are molecularly targeted to deregulated signaling pathways and the personalization of treatment planning. In oncology, personalized medicine is not just a glimmer of a future in which prevention strategies or early detection surveillance programs are tailored to DNA polymorphisms. Personalized therapeutics are already here in oncology, in terms of selecting therapy based on the genomic characterization of individual tumors. For example, treatment of women with breast cancer is often based on estrogen receptor status, *HER2* amplification status and gene-expression profiles indicating the prognostic aggressiveness of the disease.

Financial & competing interests disclosure: The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties. No writing assistance was utilized in the production of this manuscript.

Technologies such as array-based hybridization assays and next-generation DNA sequencing facilitate the identification of important molecular targets and the development of prognostic and predictive biomarkers for the personalization of therapeutic decision-making. The use of prognostic and predictive biomarkers has great potential value for cancer drug development, therapeutic decision-making for individual patients and controlling medical costs by reducing the vast over-treatment of cancer patients with therapies that do not benefit them. Nevertheless, the co-development of new drugs with companion diagnostics increases the complexity of development and may not generally provide a quicker and cheaper treatment approach, as superficial analyses often claim. Diagnostics that are not reliably evaluated can detract from proper patient management and increase the cost of medical care. One of the greatest challenges today is to develop prognostic and predictive biomarkers in a reliable, but practical, manner that permits the translation of the genomic information read from individual tumors into therapeutic strategies that benefit patients. In this review, I describe some prospective development strategies that can be of value in this process.

Biomarkers & validation

Traditionally, the term ‘biomarker’ referred to a measurement that tracks the pace of a disease, increasing as the disease progresses and decreasing as it regresses. Although there are many potential uses for such biomarkers in the early clinical development of new drugs, our focus here will be on baseline biomarkers. Prognostic markers are baseline (pretreatment) measurements that provide information about the patient's probable long-term outcome, either untreated or with a standard treatment. Prognostic markers can be used to determine whether the patient requires any systematic treatment or any therapy that is beyond the standard treatment. Predictive markers are baseline measurements that indicate whether the patient is likely (or unlikely) to benefit from a specific drug or regimen.

The medical uses of prognostic and predictive biomarkers are completely different from those of pharmacodynamic or surrogate end point biomarkers. Since ‘validation’ or ‘qualification’ only has meaning in terms of fitness for intended use, the criteria for the validation of surrogate end points should not be mistakenly applied to prognostic or predictive biomarkers [1]. The validation of prognostic and predictive biomarkers, although demanding, is often much more feasible than the validation of biomarkers as surrogate end points.

In this review, I will refer to three different types of validation for prognostic and predictive biomarkers: analytical validity, clinical validity and clinical or medical utility. When there is a gold standard measurement, analytical validity means that the test is accurate relative to the gold standard. Otherwise, analytical validity means that the test is reproducible and robust. It should be robust with regard to assay performance and tissue handling.

Clinical validity means that the test result correlates with a clinical end point or characteristic. This correlation can often be established in a retrospective study. For example, a test for predicting response to a chemotherapy regimen may be clinically validated using data from a single arm Phase II study of patients who received that regimen. Clinical validation of a test is often accompanied by calculating the sensitivity of the test for identifying responders and the specificity for identifying nonresponders. Sensitivity and specificity are computed based on a specified cut-point of positivity for the test. The receiver operating characteristic (ROC) curve is a plot of the sensitivity versus 1 minus the specificity, as a function of the cut-point. The positive and negative predictive values (NPVs) of the test depend on the sensitivity, specificity and prevalence of responders in the population in which the test will be used [2]. It is generally important to evaluate whether the test provides improved negative and positive predictive value (PPV) within the levels of standard prognostic factors. Analyses of sensitivity, specificity, ROC curves, NPV and PPV have traditionally been used with binary outcome

measures, such as tumor response. Such methods can be applied when evaluating a prognostic test for predicting the probability of patients that receive standard therapy surviving beyond a prespecified landmark time [3]. However, the methods are less suitable for evaluating whether a predictive test enables the identification of patients who will have longer survival or disease-free survival outcomes on a new treatment compared with a standard regimen.

Clinical utility means that use of the test results in improved outcome for patients. The improved outcome is generally based on using the test result to inform therapeutic decision-making. Improved outcome can mean that patients live longer, or that their disease can be managed with equivalent effectiveness and fewer adverse effects. Clinical utility requires that the test is 'actionable' and that the outcome measure really reflects patient benefit. Clinical utility requires that the clinical context and medical indication for use of the test are clear, that the magnitude of outcomes or treatment effects associated with different results of the test are sufficiently great to influence treatment decisions, and that the marker informs therapeutic decision-making that results in better patient outcome than a standard of care based on accepted prognostic factors.

I will use the term 'classifier' to refer to a diagnostic test that translates one or more biomarker measurements into a set of predicted categories. For example, with a prognostic classifier, the categories may refer to a low risk of tumor recurrence, moderate risk of recurrence and high risk of recurrence. With a predictive classifier, the categories may refer to patients who are most likely to benefit from the new regimen and those who are less likely to benefit. A gene-expression-based classifier may involve the measurement of the expression of many genes, but it is a discrete indicator of two or more classes and can be used for selecting or stratifying patients in a clinical trial, just like a classifier based on a single gene or protein. Validating a gene-expression-based classifier means evaluating whether the classifier, as a composite entity, is fit for its intended use. It does not mean determining whether the individual gene components are prognostic or predictive in a new study.

Developing gene-expression-based classifiers

In the past several years, numerous studies of gene-expression signatures as prognostic markers have been reported in the literature. The vast majority are 'developmental' studies in which prognostic expression signatures are developed based on a retrospective analysis of cases with specimens available for assay. Many of these studies suffer from a lack of focus on a defined medical use. The fact that the signatures are derived from screening thousands of genes to construct classifiers that are predictive for a small number of cases creates other potential problems, many of which are reviewed by Dupuy and Simon [4]. Although establishing the medical utility of a classifier generally requires conduction of a prospective clinical trial, the developmental study should attempt to establish the clinical validity of the classifier in the context of its intended use. The key principle to be observed when evaluating a gene-expression-based classifier is that the data used for evaluation should be distinct from the data used for developing the classifier. The developmental study should provide unbiased estimates of predictive accuracy of the new classifier within strata defined by standard prognostic factors. If the dataset is sufficiently large, separate test sets of cases that are not used for model development should be used. If the dataset is not large enough to have a separate training and test set, then complete cross-validation should be used. Complete cross-validation is a sophisticated version of training–test splits that can provide a more efficient use of limited data. Many studies use a separate training–testing split with so few cases in the test set that the results are almost meaningless [5]. However, when performed correctly, complete cross-validation is as valid as randomly splitting the data into training and test sets [6]. Unfortunately, it is often not performed correctly [4]. There is a large literature on the development of gene-

expression-based classifiers. An introduction to this literature is described by Simon *et al.* [7].

Prognostic biomarkers

The oncology literature is replete with publications on prognostic factors but very few of these are used in clinical practice [8]. Most prognostic factor studies are conducted without clear focus on a defined intended use. This lack of focus often results in the use of a heterogeneous convenience sample of patients whose tissues are available [9], and in an exploratory approach to data analysis [10] that fails to adequately address the promise of the marker for improving therapeutic decision-making.

A prognostic classifier can be therapeutically relevant if it identifies a set of patients who have a good prognosis without undergoing chemotherapy, so they may choose to be spared the risks and inconvenience of such therapy and forgo the small potential benefit. The objective of the study is not just to determine whether the marker correlates with the outcome, but whether it identifies patients with such good outcome in the absence of chemotherapy that they may choose to avoid it. Achieving these objectives requires care in the design and analysis of the study. For example, the Oncotype DX™ (Genomic Health, CA, USA) recurrence score was developed by studying women with breast cancer whose tumors were estrogen receptor positive, had not spread to the axillary lymph nodes and who had received tamoxifen as their only systemic therapy [11,12]. A score was developed, based on tumor expression of 21 genes, to identify women whose disease-free survival was sufficiently good that they might elect to forgo cytotoxic therapy. Prognostic factors developed in such a focused manner can be relevant for therapeutic decisions. The score is often used as a classifier by introducing two cut-points to distinguish patients with low, intermediate and high risks of tumor recurrence.

The analyses of many retrospective prognostic factor studies fail to adequately address the promise of the marker for enhancing therapeutic decision-making. Over-emphasis is often placed on multivariate regression modeling rather than evaluating whether the new classifier is predictive of outcome within the levels of standard prognostic factors. Odds ratios, hazards ratios and regression coefficients in multivariate analyses are not appropriate measures of predictive accuracy [4,13]. With survival or disease-free survival data, survival curves or disease-free survival curves of risk groups determined by the new classifier or time-dependent ROC curve within strata of the standard prognostic classification provide greater information concerning predictive value [3].

Ideally, a prognostic marker will be validated in a prospective clinical trial before it is ‘ready for prime time’ [14]. The marker strategy design, shown in Figure 1, is sometimes considered for evaluating the medical utility of a diagnostic test. With this design, patients are randomized to be tested or not. For those who are not tested, their treatment is determined based on stage and standard clinical prognostic factors and practice standards. For those patients who are randomized to be tested, the results of the test can be used in conjunction with stage and standard prognostic factors to inform treatment decisions. Although the marker strategy design is regarded by some as gold standard, it is often inefficient because many patients may receive the same treatment regardless of which group they are randomized to [15–17]. In order to have reasonable statistical power to detect the differences in outcome among the two randomization groups as a whole, a very large number of patients may have to be randomized. For example, suppose the end point is disease-free survival beyond 5 years and that a proportion (π) of the patients receive the same treatment regardless of which arm they are randomized to. If we want to detect a difference (Δ) in the probability of 5-year disease-free survival for patients receiving different treatments, we would have to power the study to detect a difference of $(1 - \pi)\Delta$ between the randomization groups. Since the required sample size is generally inversely proportional

to the square of the difference to be detected between the randomized groups, the required sample size will need to be very great if π is large. This inefficiency is particularly problematic for prognostic markers identifying low-risk patients for whom chemotherapy may be withheld because the prospective study is a therapeutic equivalence trial involving a small value of Δ. For example, to have 90% power (with 5% one-sided significance) for detecting a 5% point increase in the recurrence rate ($\Delta = 0.05$) from a baseline of 10%, a randomized trial of approximately 2460 low-risk patients would be required if all patients were tested and low-risk patients selected for randomization. However, the marker strategy design of Figure 1 would require approximately 9320 randomized patients to have 90% power for detecting the 2.5% point increase in recurrence rate that is only expected if a half of the patients are low risk based on the marker.

The marker strategy design may also be poorly informative in cases where the test is not just binary and the test-based treatment strategy is complex. For example, suppose that patients with a low value of the marker have chemotherapy withheld, patients with intermediate values receive standard chemotherapy and patients with high values receive intensified chemotherapy. Because the test is not performed on patients in a control group, one cannot examine results for the subsets of patients defined by test result. One is limited to just comparing the randomization groups overall.

The defects in the marker strategy design can be avoided by measuring the test in all patients and only randomizing patients for whom the treatment assignment is influenced by marker result. This modified marker strategy design, shown in Figure 2, is currently being used in the Microarray in Node-Negative and 1–3 Positive Lymph Node Disease May Avoid Chemotherapy (MINDACT) study to evaluate a 70-gene prognostic signature for determining whether to utilize chemotherapy for women with node-negative estrogen receptor-positive breast cancer [18].

The Trial Assigning Individualized Options for Treatment (TAILORx) study is a prospective clinical trial for evaluating the OncotypeDx gene-expression recurrence score for women with node-negative estrogen receptor-positive breast cancer. The main objective of the trial is to determine whether women with a low-recurrence score can have a low risk of disease recurrence even if chemotherapy is withheld. In the trial, such women are not randomized but all have chemotherapy withheld. If the recurrence score is accurate, the relapse rate for these patients would be very low and hence, the potential benefit of chemotherapy would be very small in absolute terms [19]. In the MINDACT trial, women for whom the practice standard indicates chemotherapy but who have a low risk of recurrence based on the genomic signature are randomized between a chemotherapy arm and a nonchemotherapy arm. Nevertheless, the signature will be considered to be validated if the 5-year distant metastasis-free survival rate is greater than 92% in the women randomized to having chemotherapy withheld.

By validating these prognostic signatures in a fully prospective manner, rather than by using archived tissue from a previously conducted series, one assures that an adequate number of patients are studied, that assay results are available on all patients, that the analysis is focused on a single prespecified hypothesis and that assay results reflect real-world tissue handling and laboratory variation. However, such studies are expensive and time consuming. In some cases, the effective validation of a classifier that is predictive of low-recurrence risk can be accomplished using specimens archived from an appropriate clinical trial that withheld chemotherapy from such patients. However, convincing results are only possible if the samples being analyzed are from participants of a clinical trial with a design that enables unbiased evaluation of the test, if the number of patients in the trial is sufficiently large, if the proportion with available specimens adequate for testing is high, if careful analytical validation provides assurance that assay results on archived samples are accurate predictors of assay results on

fresh tissue and if the assays are blinded to clinical data [20]. These issues are also discussed by Pepe *et al.* [21].

A prognostic biomarker can also be used to identify patients whose outcome will be very poor with standard chemotherapy. Such patients may be good candidates for experimental regimens, but unless there is a viable therapeutic option, such prognostic biomarkers may not be widely used in general practice.

Predictive biomarkers

Predictive biomarkers identify patients who are likely or unlikely to benefit from a specific treatment. For example, *HER2* amplification is a predictive classifier for benefit from trastuzumab therapy and, perhaps also, from doxorubicin [22,23] and taxanes [24]. A predictive biomarker can also be used to identify patients who are poor candidates for a particular drug; for example, advanced colorectal cancer patients whose tumors have *KRAS* mutations appear to be poor candidates for treatment with EGFR antibodies [25,26].

Predictive biomarkers may be based on single gene or protein measurements, on gene-expression classifiers, on pathway activation indicators or on disease subclassifications. Measurements based on mutation status, copy number, transcript abundance or protein expression of a single gene are often closely linked to the mechanism of action of the drug and are thus biologically interpretable. In some cases, the target of the drug is known but it is not clear how best to measure whether the target is driving tumor growth and invasion in an individual patient. For example, although trastuzumab was initially developed using a test for protein expression of *HER2*, subsequent classification has often been based on amplification of the gene [27]. In other cases, the target of the drug may not be clearly known and the options for measurement will be more numerous.

Companion diagnostics

In recognition of the molecular heterogeneity of cancer, many cancer drugs are being developed today with companion diagnostics to be used as predictive biomarkers. Sawyers has stated that:

> "One of the main barriers to further progress is identifying the biological indicators, or biomarkers, of cancer that predict who will benefit from a particular targeted therapy" [28].

This increases the complexity of drug development, although it has potential benefits for patients and for controlling medical expenses. However, it requires that an effective predictive biomarker be identified and that a test for it is analytically validated prior to the launch of the Phase III pivotal clinical trials of the drug.

The standards for evaluating the effectiveness of a new drug are well established. Ideally, one would like a randomized clinical trial that establishes that treatment of a defined selection of patients with the new drug results in improved clinical outcome over the control group. Since clinical outcome for cancer patients is usually measured by survival or progression-/recurrence-free survival, a randomized control group is usually essential. The role of the predictive biomarker for evaluating the new drug is in refining the definition of the target population.

There is considerably less clarity regarding what constitutes appropriate validation of a predictive biomarker. As described in the second section of this paper, three levels of validation can be distinguished. Analytical validation of the test for measuring the biomarker can often be based on archived tumor samples. For drugs used in regimens for treating patients with metastatic disease, clinical validation can be based on Phase II data. The evaluation involves estimating the PPVs and NPVs for identifying patients who respond to the new regimen. With

such data, ROC curves can be generated and a test cut-point selected. Regulatory and third-party payer requirements may be based on either clinical validity or medical utility, but the concept of medical utility is complex when applied to companion diagnostics for new drugs.

Medical utility generally means that use of the diagnostic results in patient benefit compared with the standard of care. If the diagnostic enables the identification of a subset of patients for whom a new drug would provide improved clinical outcome (e.g., prolonged survival) compared with a randomized standard care control group, one might argue that use of the diagnostic in conjunction with the new drug has medical utility. Attempting to dissect the medical utility of the diagnostic from the medical utility of the drug, however, may lead to intellectual and regulatory inconsistencies that could roadblock the progress in personalized and predictive medicine.

Establishing the medical utility of a companion diagnostic is generally based on the same Phase III pivotal trials that were used to establish the effectiveness of the new drug. We will take the utility of a companion diagnostic as meaning that it enables the physician to identify patients belonging to the target population who would benefit from the new drug and that test-negative patients are less likely to benefit from it. The calculation of PPV, NPV and ROC curves do not usually play a role in this analysis as they do not generally reflect the nature of the time-to-event end point and the randomized control group. In the following sections, I will review some designs for Phase III pivotal trials that utilize new drugs and companion diagnostics. Key properties of the designs are summarized in Table 1. Some of the designs incorporate both test-positive and test-negative patients and hence, are somewhat self-contained in their ability to provide evidence that the former patients benefit from a new drug and the latter patients do not. The enrichment designs described below only incorporate test-positive patients. Those designs are for situations where there is either a strong biological rationale for excluding test-negative patients from the development or compelling Phase II evidence suggesting that test-negative patients will not benefit from the new drug.

Enrichment designs

With an enrichment design, a diagnostic test is used to restrict eligibility for a randomized clinical trial comparing a regimen containing a new drug with a control regimen. This approach, shown in Figure 3, was used for the development of trastuzumab in which patients with metastatic breast cancer whose tumors expressed *HER2* in an immunohistochemistry test were eligible for randomization. Simon and Maitournam studied the efficiency of this approach relative to the standard approach of randomizing all patients without using the test at all [29–31]. They found that the efficiency of the enrichment design depended on the prevalence of test-positive patients and on the effectiveness of the new treatment in test-negative patients. When fewer than a half of the patients are test positive and the new treatment is relatively ineffective in test-negative patients, the number of randomized patients required for an enrichment design is often dramatically smaller than the number of randomized patients required for a standard design. For example, if the treatment is completely ineffective in test-negative patients, the ratio of the number of patients required for randomization in the enrichment design relative to the number required for the standard design is approximately $1/\gamma^2$, where γ denotes the proportion of patients who are test positive [31]. The treatment may have some effectiveness for test-negative patients, either because the assay is imperfect for measuring deregulation of the putative molecular target or because the drug has anti-tumor off-target effects. However, even if the new treatment is half as effective in test-negative patients as in test-positive patients, the randomization ratio will be approximately $4(\gamma + 1)^2$. This equals approximately 2.56 when $\gamma = 0.25$, that is, 25% of the patients are test positive, indicating that the enrichment design reduces the number of patients required for randomization by a factor of 2.56.

The enrichment design was very effective for the development of trastuzumab even though the immunohistochemistry assay has subsequently been replaced by a FISH-based test for *HER2* amplification. Simon and Maitournam also compared the enrichment design with the standard design with regard to the number of screened patients [29–31]. Zhao and Simon have made the methods for sample size planning in the design of enrichment trials available online at [101]. The web-based programs are available for binary and survival/disease-free survival end points. The planning takes into account the performance characteristics of the tests and the specificity of the treatment effects. The programs provide comparisons with standard nonenrichment designs based on the number of randomized patients required and the number of patients needed for screening to obtain the required number of randomized patients.

The enrichment design is particularly appropriate for contexts in which there is a strong biological basis for believing that test-negative patients will not benefit from the new drug and that administering the drug in these patients would raise ethical concerns. In many situations, the biological basis is strong but not compelling. The enrichment design does not provide data regarding the effectiveness of the new treatment compared with a control for test-negative patients. Consequently, unless there is Phase II data on the clinical validity of the test for predicting response, or compelling biological evidence that the new drug is not effective in test-negative patients, the enrichment design may not be adequate to support approval of the test.

Designs that include both test-positive & test-negative patients

When a predictive classifier has been developed but there is no compelling biological or Phase II data suggesting that test-negative patients do not benefit from the new treatment, it is generally best to include both classifier-positive and classifier-negative patients in the Phase III clinical trials comparing the new treatment with the control regimen (Figure 4). In this case, it is essential that an analysis plan is predefined in the protocol explaining how the predictive classifier will be used in the analysis. It is not sufficient just to stratify, that is, to balance the randomization with regard to the classifier without specifying a complete analysis plan. The main value of ‘stratifying’ (i.e., balancing) the randomization is to ensure that only patients with adequate test results will enter the trial. Prestratification of the randomization is not necessary for the validity of inferences to be made regarding treatment effects within the test-positive or test-negative subsets. If an analytically validated test is not available at the start of the trial but will be available by the time of analysis, it may be preferable not to prestratify the randomization process. Similarly, if the predictive biomarker to be used in the analysis is not completely settled at the start of the trial but will be determined based on external data by the time of analysis, careful prespecification of the analysis plan will be necessary, but prestratification of the randomization process will not be appropriate.

The purpose of the pivotal trial is to evaluate the new treatment, both overall and in the subsets determined by the prespecified classifier. The purpose is not to modify or optimize the classifier. If the classifier is a composite gene-expression-based classifier, the purpose of the design is not to re-examine the contributions of each gene. If one examines any of this, then an additional Phase III trial might be needed to evaluate the treatment benefit in subsets determined by the new classifier. Several primary analysis plans are presented below to illustrate that the plan should stipulate in detail how the predictive biomarker will be used in the analysis and that there should be no exploratory aspect to the treatment evaluation. These strategies are discussed in greater detail by Simon [32,33], and a web-based tool for sample size planning with these analysis plans is available at [101].

Analysis plan for a biomarker with strong credentials—If one does not expect the treatment to be effective in the test-negative patients unless it is effective in the test-positive

patients, one might first compare the treatment versus a control in test-positive patients using a threshold of significance of 5%. The new treatment will only be compared with the control among test-negative patients, again using a threshold of statistical significance of 5%, if the treatment versus control comparison is significant at the 5% level in test-positive patients. This sequential approach controls the overall type I error at 5%.

To have 90% power in the test-positive patients for detecting a 50% reduction in hazard for the new treatment versus control at a two-sided 5% significance level, approximately 88 events of test-positive patients are required. If, at the time of analysis, the event rates in the test-positive and test-negative strata are approximately equal, when there are 88 events in the test-positive patients there will be approximately $88(1 - \gamma)/\gamma$ events in the test-negative patients, where γ denotes the proportion of test-positive patients. If 25% of the patients are test positive, there will be approximately 264 events in test-negative patients. This will provide approximately 90% power for detecting a 33% reduction in hazard at a two-sided significance level of 5%. In this case, the trial will not be delayed compared with the enrichment design, but a large number of test-negative patients will be randomized, treated and followed on the study rather than excluded, as with the enrichment design. This will be problematic if one does not, *a priori*, expect the new treatment to be effective for test-negative patients. In this case, it will be important to establish an interim monitoring plan to terminate accrual of test-negative patients when the interim results and prior evidence of lack of effectiveness make it no longer viable to include these patients.

Fall-back analysis plan—In the situation where one has limited confidence in the predictive marker, it can be effectively used for a ‘fall-back’ analysis. Simon and Wang proposed an analysis plan in which the new treatment group is first compared with the control group overall [34]. If that difference is not significant at a reduced significance level, such as 0.03, the new treatment is compared with the control group for test-positive patients only. The latter comparison uses a threshold of significance of 0.02, or whatever portion of the traditional 0.05 is not used by the initial test.

If the trial is planned for having 90% power for detecting a uniform 33% reduction in overall hazard using a two-sided significance level of 0.03, the overall analysis will take place when there are 297 events. If the test is positive in 25% of patients and the event rates in test-positive and test-negative patients are equal at the time of analysis, when there are 297 overall events there will be approximately 75 events among the test-positive patients. If the overall test of treatment effect is not significant, the subset test will have a 75% power for detecting a 50% reduction in hazard at a two-sided 0.02 significance level. By delaying the treatment evaluation in the test-positive patients, 80% power can be achieved when there are 84 events and a power of 90% can be achieved when there are 109 events in the test-positive subset.

Wang *et al*. have demonstrated that the power of this approach can be improved by taking into account the correlation between the overall significance test and the significance test comparing treatment groups in the subset of test-positive patients [35]. So if, for example, a significance threshold of 0.03 has been used for the overall test, the significance threshold used for the subset can be somewhat greater than 0.02 and the overall chance of any type of a false-positive claim occuring will be limited to 5%.

Interaction analysis plan—A third possible analysis plan is to decide whether to compare the treatments overall or within the test-positive and test-negative subsets based on a preliminary test of interaction. The interaction test comparing the treatment effects for those two subsets should be one-sided and performed at a threshold above the traditional 5% level. Testing for interaction is generally not really the purpose of the study and can require substantially more patients than for comparing the new treatment with the control in test-

positive and test-negative subsets. In the example described above, the interaction test will have approximately 93.7% power at a one-sided significance level of 0.10 for detecting an interaction with a 50% reduction in hazard for test-positive patients and no treatment effect in test-negative patients. Detailed results of this analysis plan are available using the web-based program described above.

Adaptive clinical trial designs using predictive biomarkers

Adaptively modifying types of patients accrued

Wang *et al.* proposed a Phase III design comparing a new treatment with a control, which starts with accruing both test-positive and test-negative patients [35]. An interim analysis is performed, evaluating the new treatment in the test-negative patients. If the observed efficacy for the control group exceeds that of the new treatment group and the difference exceeds a futility boundary, then accrual is limited to test-positive patients until the originally planned total sample size is reached. Wang *et al.* demonstrated computer simulations indicating that this design has greater statistical power than nonadaptive approaches, but their design involves many more test-positive patients and may require a much longer trial duration. The concept of curtailing accrual of test-negative patients based on an interim futility analysis could be implemented, but without the extension of trial duration resulting from the substitution of test-positive for test-negative patients to achieve the prespecified total sample size. Nevertheless, a futility analysis for test-negative patients is only likely to be effective if the time to observing a patient's end point is rapid relative to the accrual rate.

Liu *et al.* proposed a two-stage design [36] in which only test-positive patients are accrued during the initial stage of the trial. At the end of the first stage, an interim analysis is performed comparing the outcome of the new treatment versus the control for the test-positive patients. If the results are not promising for the new treatment, accrual stops and no treatment benefit is claimed. If the results are promising for the test-positive patients at the end of the first stage, accrual continues for test-positive patients and accrual also commences for test-negative patients in the second stage.

Adaptive threshold design

Jiang *et al.* reported on a ‘Biomarker Adaptive Threshold Design’ for situations where a specific predictive index or biomarker score is available at the start of the trial, but a cut-point for converting the score into a binary classifier is not established [37]. With their design, tumor specimens are collected from all patients at trial entry, but the value of the predictive index is not used as an eligibility criteria. Their analysis plan does not stipulate that the assay for measuring the index needs to be performed in real time, although such stratification could be employed. Jiang *et al.* described two analysis plans. Analysis plan A begins with comparing the outcomes for all patients receiving the new treatment with those for all control patients. If this difference in outcomes is significant at a prespecified significance level (α_1), the new treatment is considered effective for the eligible population as a whole. Otherwise, a second-stage test is performed using the significance threshold of $\alpha_2 = 0.05\text{-}\alpha_1$. The second-stage test involves finding the cut-point b* for which the difference in outcome of the treatment versus control (i.e., the treatment effect) is maximized when the comparison is restricted to patients with predictive index scores above that cut-point. The statistical significance of that maximized treatment effect is determined by generating the null distribution of the maximized treatment effect under random permutations of the treatment labels. If the maximized treatment effect is significant at level α_2 of this null distribution, the test treatment is considered effective for the subset of patients with a biomarker value above the cut-point at which the maximum treatment effect occurred. Jiang *et al.* described construction of a confidence interval for the optimal cut-point using a bootstrap resampling approach and described sample size planning for this design.

Adaptive biomarker design

For the co-development of a new drug and companion diagnostic, it is best to have the candidate diagnostic completely specified and analytically validated prior to its use in the pivotal clinical trials. However, this is difficult and in some cases is not feasible. The approach of Jiang *et al.* [37] can be generalized to the setting in which one has several candidate predictive classifiers: B_1, B_2 … B_K. Let S(k) denote the log-likelihood measure of treatment effect for patients who are positive for biomarker B_k and let k* denote the biomarker for which S(k) is maximum. The statistical significance of S(k*) is determined by permuting the treatment group labels of the patients and then re-evaluating the treatment effects within the positive subsets of the K binary classifiers. Using bootstrap resampling, one can evaluate the proportion of the times that each patient is included in the positive subset of the selected biomarker and obtain a confidence interval for the treatment effect in the selected subset.

Adaptive signature design

Freidlin and Simon proposed a design for a Phase III trial that can be used when no classifier is available at the start of the trial [38]. The design provides for the development of the classifier and the evaluation of treatment effects in a single trial while preserving the principle of separating the data used for developing a classifier from the data used for evaluating treatment in subsets determined by the classifier.

At the conclusion of the trial, the new treatment is compared with the control overall, using a threshold of significance of α_1, which is somewhat less than the total α. A finding of statistical significance at that level is taken as support of a claim that the treatment is broadly effective. At that point, no biomarkers have been tested on the patients, although patients must have tumor specimens collected to be eligible for the clinical trial.

If the overall treatment effect is not significant at the α_1 level, a second stage of analysis takes place. The patients are divided into a training set and a testing set. The data for patients in the training set is used to define a single subset of patients who are expected to be most likely to benefit from the new treatment compared with the control. Freidlin and Simon used a machine learning algorithm based on screening thousands of genes for those with expression values that interact with the treatment effect, but the design can be used with other algorithms and even with candidate classifiers that do not involve gene expression. When that subset has been explicitly defined, the new treatment is compared with the control for patients in the test set who display the characteristics defined by that subset. The comparison of the new treatment with the control in the subset is restricted to patients in the test set in order to preserve the principle of separating the data used to develop a classifier from the data used to test treatment effects in subsets defined by that classifier. The comparison of treatment with control for the subset uses a threshold of significance of $\alpha - \alpha_1$ in order to ensure that the overall chance of a false-positive conclusion is no greater than α. These thresholds can be sharpened using the methods of Song and Chi [39].

Friedlin and Simon demonstrated that the adaptive signature design can be effective for the development and use of gene-expression classifiers if there is a very large treatment effect in a subset determined by a set of signature genes. However, the power of the procedure for identifying the subset is limited by having to test the treatment effect at a stringent significance level in the subset of patients who are restricted to the test set that is not used for classifier development. The analysis strategy used by the adaptive signature design can be used more broadly than just in the context of identifying *de novo* gene-expression signatures. For example, it could be used when several gene-expression signatures are available at the outset and it is not clear which to include in the final statistical testing plan. It could also be used with classifiers based on a single gene but with several candidate tests for measuring expression or deregulation

of that gene. For example, while the focus may be on EGFR, there may be uncertainty regarding whether to measure overexpression at the protein level, point mutation of the gene or amplification of the gene. In these settings with a few candidate classifiers, a smaller training set may suffice instead of the 50:50 split used by Freidlin and Simon.

Conclusion

Developments in cancer genomics and biotechnology are changing the opportunities for the development of more effective therapeutics and molecular diagnostics to guide the use of those drugs. These opportunities can have enormous potential benefits for patients and for containing healthcare costs. One of the greatest opportunities is in developing predictive biomarkers for patients who require treatment and are (or are not) likely to benefit from specific drugs.

However, co-development of drugs and companion diagnostics adds complexity to the development process. Traditional *post hoc* correlative science paradigms do not provide an adequate basis for reliable predictive medicine. New paradigms are required for separating biomarker development from therapeutic evaluation. Without rigorous validation based on the intended use of these developments, oncology could be inundated with expensive tests of uncertain medical utility. New clinical trial designs are required that incorporate prospective analysis plans that provide flexibility in identifying the appropriate target population in a manner that preserves overall false-positive error rates. Such analysis plans must be constructed to provide information regarding the specificity of treatment effects without requiring sample sizes that are great enough to discourage the development of predictive biomarkers or to require physicians to expose large numbers of patients to drugs from which they are not expected to benefit.

New regulatory approaches to the evaluation of drugs and companion diagnostics are also necessary. Medical utility means that use of the new drug with the diagnostic results in increased patient benefit compared with the standard of care. If the diagnostic enables the identification of a subset of patients for whom a new drug provides improved clinical outcome (e.g., prolongs survival) compared with a randomized standard care control group, one might argue that the use of the diagnostic in conjunction with the new drug has medical utility.

Described above are some of the new prospective approaches that are available for the reliable evaluation of prognostic and predictive biomarkers. These approaches include targeted enrichment designs for settings where biological evidence or Phase II data destroy the equipoise necessary to include test-negative patients in the Phase III clinical trial. We have emphasized that for designs that do not use the predictive biomarker as an exclusion criterion, it is essential to have a specific prospectively-defined analysis plan outlining exactly how the new treatment will be evaluated with regard to the test. Because of the complexity of the biology of chronic diseases such as cancer, it is not always feasible to identify a single appropriate candidate predictive biomarker and to develop an analytically validated test by the initiation of Phase III pivotal trials of a new drug. We have described several carefully controlled adaptive designs for utilizing the trial data in order to refine the biomarker and provide valid Phase III-level analyses of treatment effects.

Adapting to the fundamental heterogeneity of many human diseases and achieving the benefits of personalized predictive medicine for patients and for the economics of healthcare will require paradigm changes for academic clinical investigation, industry drug development and regulatory evaluation. I have attempted to identify some of the key issues involved and to provide some guidance on the design of clinical trials for evaluating the clinical utility of prognostic and predictive biomarkers.

Future perspective

Personalized medicine is already established in many parts of oncology today and is progressing rapidly. Developments in oncology provide an indication of the direction of personalized medicine in other diseases. Personalized medicine in oncology is not based on intervention strategies for patients at high risk of disease owing to inherited genetics, it is based on the classification of individual tumors based on the somatic mutations that they contain. Although tumors of a given primary site may be highly individualized in terms of their mutational spectrum, a limited number of key pathways are common targets. It appears likely that the tumors of a primary site will be effectively classifiable into a limited number of subtypes based on the pathways that have been deregulated during oncogenesis. Most oncologic therapeutic development will be based on these deregulated pathways and clinical trials will be increasingly evaluated with regard to pathway-based disease classifications assessed by mutational analysis. Because of the complexity of signaling pathways and the difficulty of elucidating the steps of oncogenesis, the move to a fully selective mode of drug development will occur over a decade, but it has already gathered momentum. Improved disease classification and mutational evaluation of individual tumors will make the empirical development of drug-specific predictive biomarkers less burdensome. Clinical validation of predictive biomarkers for use with specific drugs based on Phase II studies will still be important, but Phase III studies will be increasingly conducted in molecularly-targeted enriched populations. The treatment of patients in general practice will increasingly be based on disease classifications driven by pathway deregulation and assessed by DNA sequencing of tumor samples.

Executive summary

- A prognostic biomarker is a baseline measurement that provides information on patient outcome, either untreated or with a defined standard treatment. Prognostic biomarkers are most likely to be therapeutically relevant if they are developed with an intended medical use clearly in mind. Two common intended uses are:
  - Identifying low-stage patients who have a poor prognosis with the minimally invasive treatment, which is the practice standard for their stage;
  - Identifying patients who have a very good prognosis with a minimally invasive treatment that is no longer the practice standard for their stage.
- A predictive biomarker should identify patients who are likely (or unlikely) to benefit from a specific treatment.
- Traditional exploratory analyses of randomized clinical trials may be useful for predictive biomarker discovery, but they do not provide an adequate framework for establishing the fitness of the marker for broad application.
- Co-development of therapeutics and companion diagnostics increases the complexity of all stages of the development process. Nevertheless, companion diagnostics can have great value for patients, for controlling healthcare costs and for improving the chance of successful drug development.
- Companion diagnostics are particularly important for the effective development of a therapeutic in cases where fewer than a half of the conventionally diagnosed patients are likely to benefit from the drug.

- The standards for evaluating effectiveness of a new drug are well established, generally involving a randomized clinical trial, which establishes that treatment of a defined target population of patients with the new drug results in improved clinical outcome compared with the control group. The role of the predictive biomarker for evaluating the new drug is in refining the definition of the target population. There is considerably less clarity regarding what constitutes appropriate 'validation' of a predictive biomarker.
- Three levels of validation can be distinguished. Analytical validation of the test for measuring the biomarker can often be based on archived tumor samples. For cancer drugs used as single agents for treating patients with metastatic disease, clinical validation can be based on single arm Phase II data. The evaluation involves estimating the positive and negative predictive values for identifying patients who respond to the new regimen. With such data, receiver operating characteristic curves can be generated and a test cut-point selected. For establishing the medical utility of a predictive biomarker, randomized clinical trials are generally required that demonstrate that the marker distinguishes a subset of patients who benefit from the regimen from those who do not. Traditional concepts of positive predictive value, negative predictive value and receiver operating characteristic curves do not directly apply.
- New regulatory approaches to the evaluation of drugs and companion diagnostics are necessary. Medical utility means that use of the new drug with the diagnostic results in patient benefit compared with the standard of care. If the diagnostic enables the identification of a subset of patients for whom a new drug provides improved clinical outcome, then use of the diagnostic in conjunction with the new drug has medical utility. Attempting to dissect the medical utility of the diagnostic from the medical utility of the drug may lead to intellectual and regulatory inconsistencies that could roadblock progress in personalized and predictive medicine.
- For pivotal trials of a new drug utilizing a companion diagnostic, it is desirable to have a completely specified analytically validated test available at the start of the study.
- When there is a compelling biological basis for expecting that test-negative patients are unlikely to benefit from the new drug, they may be excluded from the pivotal trial using an 'enrichment design'. This may lead to a very efficient randomized clinical trial for evaluating the new treatment. In some cases, Phase II data may provide strong evidence that test-negative patients are unlikely to benefit from the new drug and can be used to establish the clinical validity and medical utility of the test.
- In cases where the biological basis for only selecting test-positive patients for study is less than compelling, both test-positive and test-negative patients should be included in the randomized pivotal trial.
  - Stratification of the randomization by the test helps to ensure that diagnostic specimens are available and test results are obtained for all randomized patients. The analysis strategy for the pivotal trial should be completely specified prospectively and several analysis strategies are described in this review.
  - With a detailed prospective plan for the primary analysis that preserves the study-wise type I error, a claim for treatment benefit for the test-positive subset should not be contingent on establishing that the treatment

is effective for the overall population, nor on establishing that there is a treatment by subset interaction that is significant at the traditional 5% level.

- Adaptively refining the predictive biomarker using the methods described by Jiang *et al.* [37] and by Freidlin and Simon [38] provides flexibility to the pivotal trial while preserving statistical rigor. These methods can be generalized to evaluating treatment effects for the biomarker-positive cases using multiple prespecified candidate binary biomarkers and are not restricted to using gene-expression signature biomarkers.
- Adapting to the fundamental heterogeneity of many human diseases and achieving the benefits of personalized predictive medicine for patients and for the economics of healthcare will require paradigm changes for academic clinical investigation, industry drug development and regulatory evaluation.

Bibliography

Papers of special note have been highlighted as:

▪ of interest

▪▪ of considerable interest

1. Chau CH, Rixe O, McLeod H, Figg WD. Validation of analytical methods for biomarkers employed in drug development. Clin Cancer Res 2008;14(19):5967–5976. [PubMed: 18829475]
2. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; Oxford, UK: 2004.
3. Heagerty PJ, Lumley T, Pepe MS. Time dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2004;56:337–344. [PubMed: 10877287]
4. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147–157. [PubMed: 17227998]
5. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–18. [PubMed: 12509396]
6. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21(15):3301–3307. [PubMed: 15905277]
7. Simon, R.; Korn, EL.; McShane, LM.; Radmacher, MD.; Wright, GW.; Zhao, Y. Design and Analysis of DNA Microarray Investigations. Springer Verlag; NY, USA: 2003.
8. Pusztai L. Perspectives and challenges of clinical pharmacogenomics in cancer. Pharmacogenomics 2004;5(5):451–454. [PubMed: 15212580]
9. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–985. [PubMed: 8198989]
10. Henry NL, Hayes DF. Uses and abuses of tumor markers in the diagnosis, monitoring and treatment of primary and metastatic breast cancer. Oncologist 2006;11:541–552. [PubMed: 16794234]
11. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–2826. [PubMed: 15591335] ▪ Provides a good case study of the development of an effective prognostic biomarker.
12. Paik S. Development and clinical utility of a 21-gene recurrence score prognostic assay in patients with early breast cancer treated with tamoxifen. Oncologist 2007;12:631–635. [PubMed: 17602054]
13. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in guaging the performance of a diagnostic, prognostic or screening marker. Am J Epidemiol 2004;159:882–890. [PubMed: 15105181] ▪ Important demonstration of the distinction between measures of association and measures of predictive accuracy.

14. Simon R. When is a genomic classifier ready for prime time? Nat Clin Pract Oncol 2004;1(1):2–3. [PubMed: 16264773]
15. Mandrekar S, Grothey A, Goetz M, Sargent D. Clinical trial designs for prospective validation of biomarkers. Am J Pharmacogenomics 2005;5(5):317–325. [PubMed: 16196501]
16. Hoering A, Leblanc M, Crowley J. Randomized Phase III clinical trial designs for targeted agents. Clin Cancer Res 2008;14(14):4358–4367. [PubMed: 18628448]
17. Mandrekar S, Sargent D. Clinical trial designs for predictive biomarker validation: one size does not fit all. J Biopharm Stat 2009;19:530–542. [PubMed: 19384694]
18. Bogaerts J, Cardoso F, Buyse M, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 2006;3(10):540–551. [PubMed: 17019432] ▪ Good review of the design considerations for a prospective trial to validate a prognostic biomarker.
19. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol 2008;26(5):721–728. [PubMed: 18258979]
20. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. Natl Cancer Inst 2009;101(21):1446–1452.
21. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter J. Pivotal evaluation of the accuracy of a classification biomarker: the probe study design. J Natl Cancer Inst 2008;10:0, 432–438.▪ Important paper on an improved approach to the design of studies for the evaluation of biomarkers used for early disease detection.
22. Hayes DF. Prognostic and predictive factors revisited. Breast 2005;14:493–499. [PubMed: 16239111]
23. Gennari A, Sormani MP, Pronzato P, et al. *HER2* status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized clinical trials. J Natl Cancer Inst 2008;100(1): 14–20. [PubMed: 18159072]
24. Hayes DF, Thor AD, Dressler LG, et al. HER2 and response to paclitaxel in node-positive breast cancer. N Engl J Med 2007;357:1496–1506. [PubMed: 17928597]
25. Amado RG, Wolf M, Peeters M, et al. Wild-type *KRAS* is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 2008;26(10):1626–1634. [PubMed: 18316791]
26. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. *K-ras* mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 2008;359(17):1757–1765. [PubMed: 18946061]
27. Wolff AC, Hammond EH, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 2007;25(1):118–145. [PubMed: 17159189]
28. Sawyers CL. The cancer biomarker problem. Nature 2008;452:548–552. [PubMed: 18385728] ▪▪ Very useful review of the importance and challenges in the development of predictive biomarkers.
29. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 2005;10:6759–6763. [PubMed: 15501951] ▪▪ Quantitative evaluation of the efficiency of targeted enrichment designs.
30. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and correction. Clin Cancer Res 2006;12:3229.
31. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Stat Med 2005;24:329–339. [PubMed: 15551403]
32. Simon R. Using genomics in clinical trial design. Clin Cancer Res 2008;14:5984–5993. [PubMed: 18829477] ▪▪ Provides a roadmap for the prospective validation of predictive biomarkers.
33. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. Expert Rev Mol Diagn 2008;2(6):721–729.
34. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. Pharmacogenomics J 2006;6:1667–1173.
35. Wang SJ, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat 2007;6:227–244. [PubMed: 17688238]
36. Liu A, Li Q, Yu KF, Yuan VW. A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. Stat Med. 2009 In press.

37. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst 2007;99:1036–1043. [PubMed: 17596577]

38. 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;11:7872–7878. [PubMed: 16278411]

39. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. Stat Med 2007;26:3535–3549. [PubMed: 17266164]

Website

101. National Cancer Institute: Biometric Research Branch. http://brb.nci.nih.gov

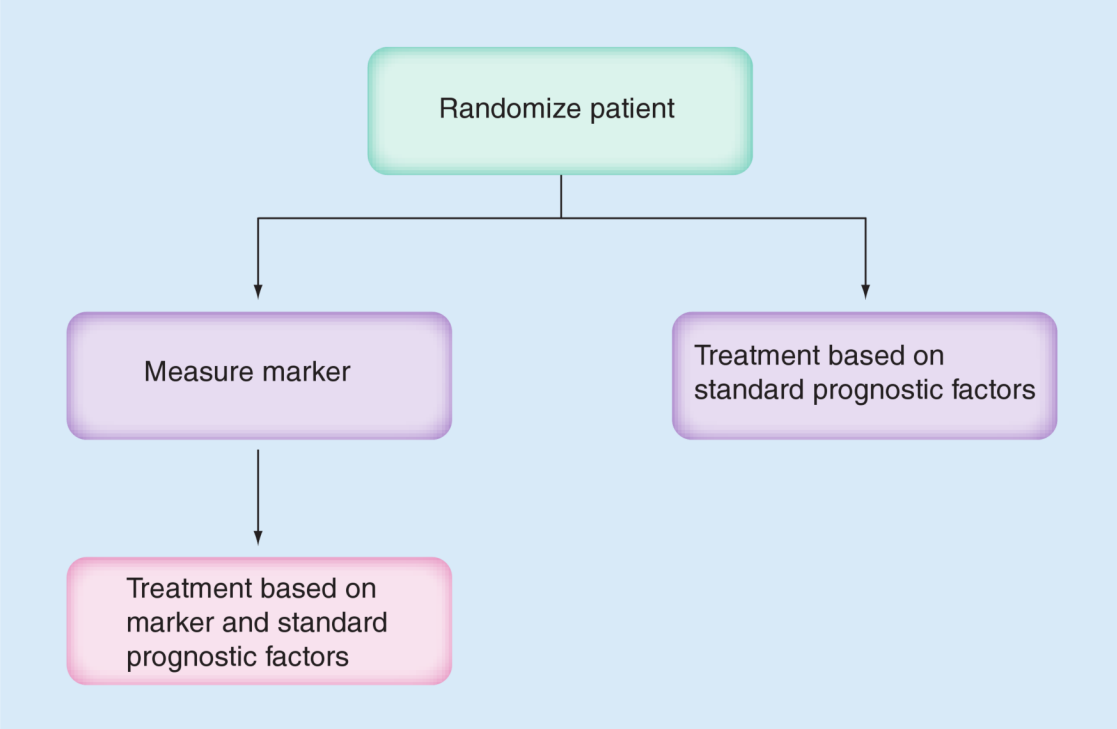

Figure 1. Marker strategy design randomizes eligible patients between two treatment assignment strategies

The control arm determines treatment using practice standards based on staging and existing prognostic factors. The new biomarker is not measured for patients that are randomized to the control arm. Patients randomized to the experimental arm have the candidate biomarker measured and this is used in conjunction with staging and other prognostic factors to determine treatment. This design is very flexible, but often very inefficient in the sense that the same objectives can be obtained with fewer patients using other designs.

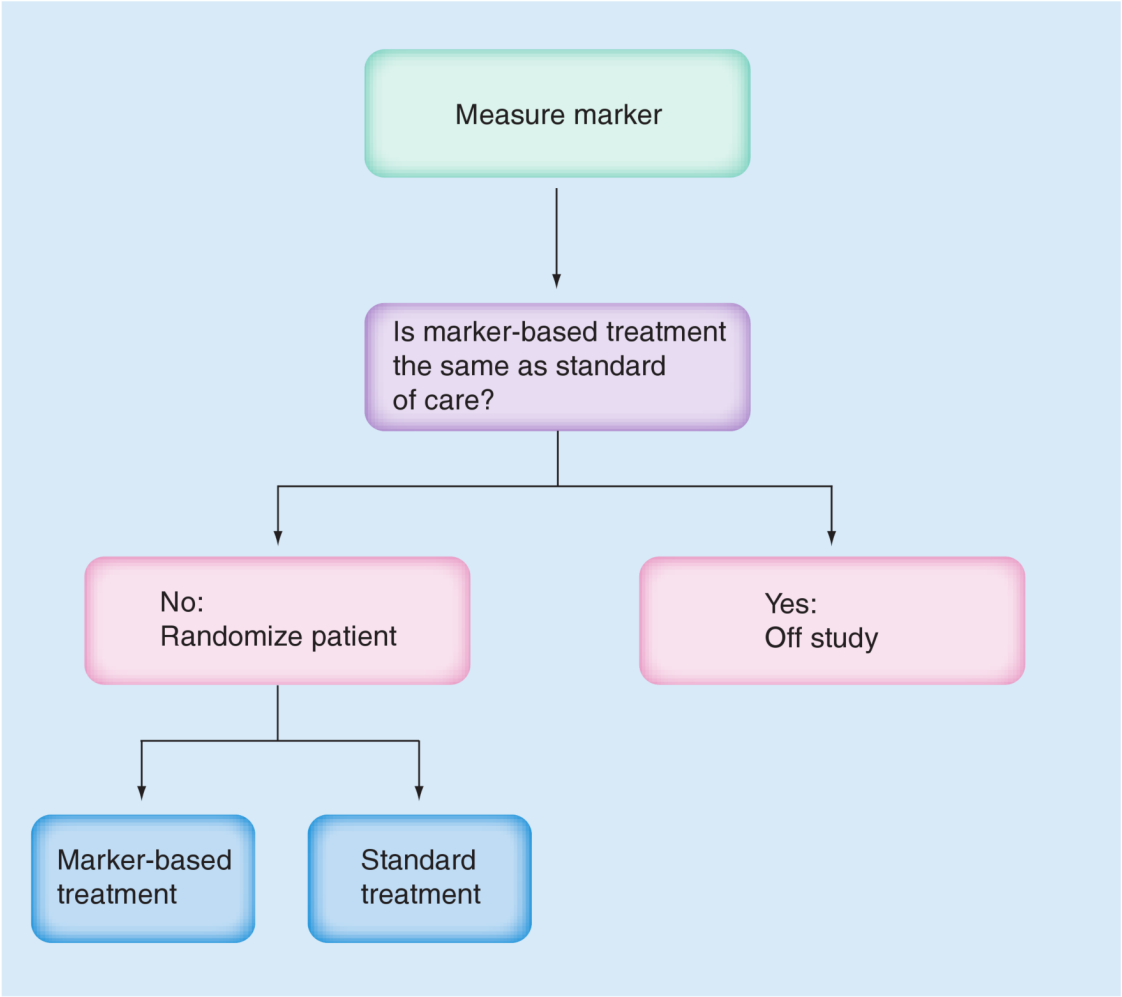

Figure 2. Modified marker strategy design measures the candidate marker in all eligible patients
Before randomization, the practice standard-determined treatment and the marker-based treatment are identified. Only patients for whom the two treatments differ are randomized. This design is generally much more efficient than the marker strategy design.

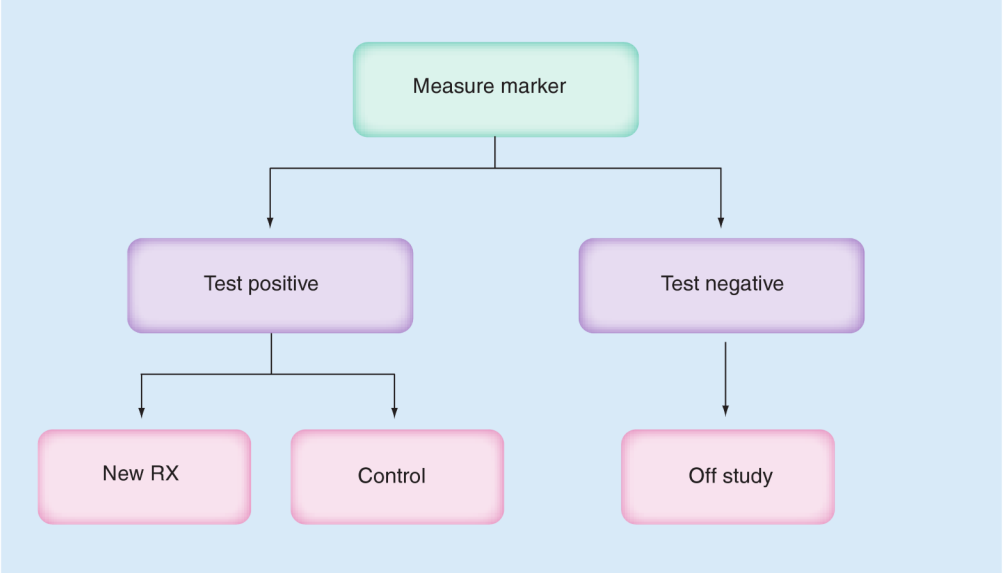

Figure 3. Targeted enrichment design is used for evaluating a new treatment in the population of patients who are identified using a predictive biomarker as the best candidates to receive potential benefit from the new treatment

The targeted enrichment design is primarily for settings where there is a compelling basis for not expecting that marker-negative patients can benefit from the new treatment and an analytically accurate test is available. The compelling basis is generally based on biology but could be based on substantial prior evidence for the new treatment. When the proportion of marker-positive patients is less than a half, this design can require substantially fewer randomized patients than the standard design. RX: Treatment.

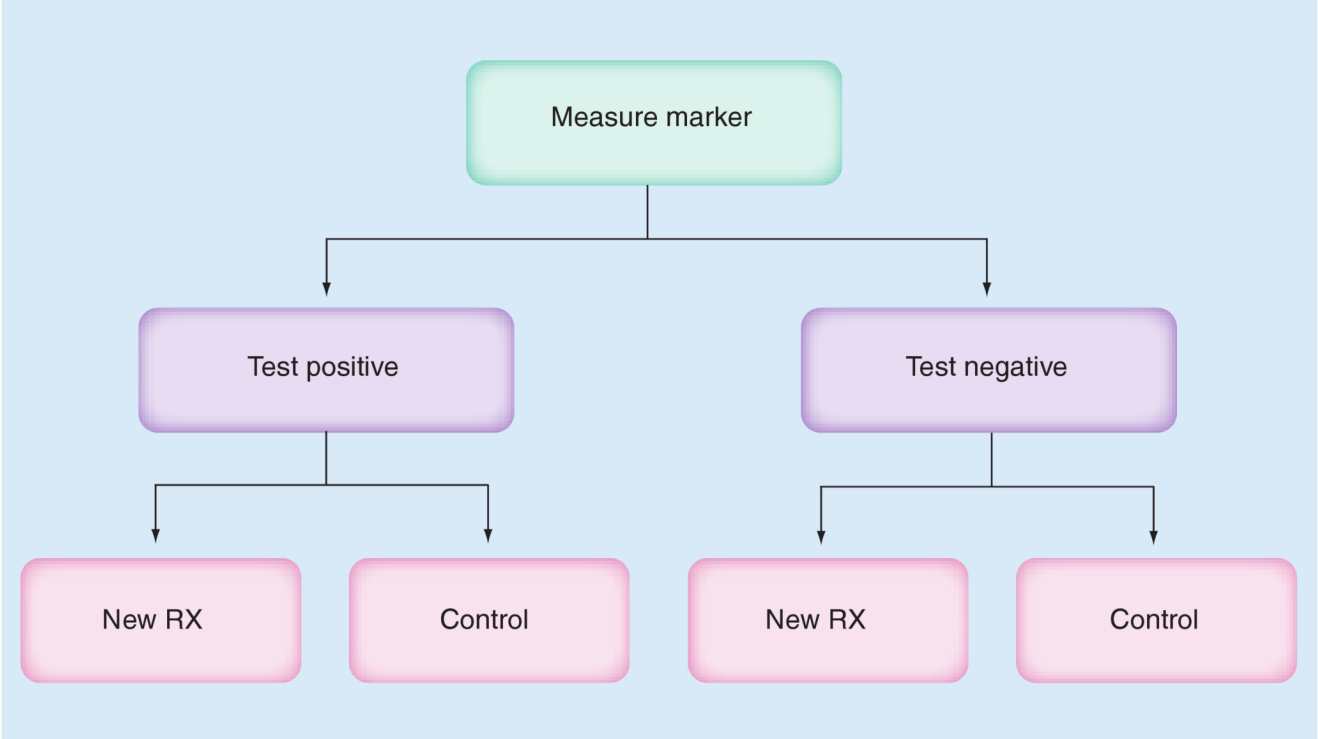

Figure 4. The stratification design is used for evaluating the effectiveness of a new treatment versus a control in a population that is prospectively characterized by a binary predictive biomarker

A detailed prospective plan should describe the primary comparison of the treatment with the control overall and in the marker-positive and marker-negative subsets. Several analysis plans are described in the text. With a focused analysis plan, claims of treatment effectiveness in marker-positive patients need not be restricted to cases where the treatment is effective overall for all patients. Ideally, a single completely defined, analytically defined binary biomarker will be determined prior to the randomized trial. Adaptive modifications of the stratified design in which the biomarker is refined based on trial data are described in the text. RX: Treatment.

Table 1
Designs for a Phase III pivotal trial of a new treatment with a predictive biomarker

Design	When to use	Strengths	Limitations
Enrichment	When strong biological evidence suggests that potential treatment effectiveness is limited to test positives	Small number of randomized patients required	Does not provide data for establishing that treatment is ineffective for test negatives
Include test negatives and positives Analysis plans illustrated for: • Strong confidence in a biomarker • Fall-back analysis of test positives • Preliminary interaction test	When enrichment design is not appropriate but a single predictive biomarker and cut-point have been determined	Permits establishing utility of the treatment and test	Requires a single biomarker and cut-point to be defined in advanceRequires a sample size large enough to evaluate treatment in test negatives and test positives separately
Adaptive threshold design (single biomarker without cut-point defined in advance)	When the threshold for the positivity of the test is not established at the start of a pivotal trial	Permits establishing utility of the treatment and testReduces dependence on Phase II data for establishing a test cut-point	Requires a single biomarker (but not cut-point) to be defined in advance
Adaptive biomarker design (multiple binary biomarkers defined in advance)	When there are a small number of candidate binary biomarkers defined in advance	Does not require that a single biomarker be defined in advance	Requires that each biomarker has a defined cut-pointIncreased sample size required
Adaptive signature design: Develops a predictive signature in a training set of the trial and evaluates the treatment effect for signature and patients in the test set	When emphasis is on overall treatment effect but a fall-back secondary analysis is desired and no biomarker is available	Enables the test to be determined based on randomized data for patients included in the pivotal trial	Limited power for testing treatment effectiveness in the signature-positive subsetRequires expression profiling of trial patients
Generalized adaptive signature design: Uses the training set of the trial to select among candidate biomarkers and to optimize cut-points; the selected biomarker is evaluated in the test set	When several candidate biomarkers are available but Phase III data is needed to refine and select among them	Enables the test to be optimized based on randomized data for patients included in the pivotal trial	Limited power for testing treatment effectiveness in the test-positive subset