Expert Opinion

Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics

Richard Simon
National Cancer Institute, Biometric Research Branch, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA

Background: Developments in genomics and biotechnology provide unprecedented opportunities for the development of effective therapeutics and companion diagnostics for matching the right drug to the right patient. Effective co-development involves many new challenges with increased opportunity for success as well as delay and failure. *Objective*: Clinical trial designs and adaptive analysis plans for the prospective design of pivotal trials of new therapeutics and companion diagnostics are reviewed. *Conclusions*: Effective co-development requires careful prospective planning of the design and analysis strategy for pivotal clinical trials. Randomized clinical trials continue to be important for evaluating the effectiveness of new treatments, but the target populations for analysis should be prospectively specified based on the companion diagnostic. *Post hoc* analyses of traditionally designed randomized clinical trials are often deeply problematic. Clear separation is generally required of the data used for developing the diagnostic test, including their threshold of positivity, from the data used for evaluating treatment effectiveness in subsets determined by the test. Adaptive analysis can be used to provide flexibility to the analysis but the use of such methods requires careful planning and prospective definition in order to assure that the pivotal trial adequately limits the chance of erroneous conclusions.

Keywords: biostatistics, clinical trial design, companion diagnostics, predictive biomarker

Expert Opin. Med. Diagn. (2008) 2(6):721-729

1. Introduction

Clinical trials of new drugs have traditionally been conducted with broad patient populations. Broad eligibility criteria have generally been encouraged to avoid discrepancies between the population tested and the population eventually treated with the drug. In oncology, however, this has resulted in treating many for the benefit of the few. For example, only ~ 5% of women with estrogen receptor positive breast cancer that has not spread to the axilla benefit from cytotoxic chemotherapy. For prevention studies, the number treated to benefit one patient is even more extreme. This over-treatment results in a substantial number of adverse events and expense for treatment of patients who receive no benefit. In oncology, accumulating understanding of genomic differences among tumors of the same primary site indicates that most molecularly targeted agents are likely to benefit only the patients whose tumors are driven by deregulation of the targeted pathways [1]. Availability of improved tools for characterizing tumors biologically makes it increasingly possible to predict whether the tumor will be responsive to a particular treatment [2]. It is crucial that new drugs be developed

informa
healthcare

with companion diagnostics that identify the patients who are good candidates for treatment. It is often very difficult to perform adequate studies that identify which patients are good candidates for treatment after the treatment has been approved and used broadly. Successful prospective co-development of a drug and companion diagnostic presents many new challenges, however. In this paper, some of the issues in the use of adaptive designs of Phase III clinical trials for new treatments and diagnostic tests that may indicate which patients benefit or do not benefit from the new treatment are addressed.

2. Prognostic and predictive classifiers

The medical literature is replete with publications on prognostic factors, but very few of these are used in clinical practice. For example, Pusztai *et al.* [3] identified 939 publications over a 20-year period on prognostic factors in breast cancer, but only 4 factors (ER, PR, HER2 and Oncotype DX) are recommended for use by the American Society of Clinical Oncology. Prognostic factors are rarely used unless they help with therapeutic decision-making. Most prognostic factor studies are conducted using a convenience sample of patients whose tissues are available [4]. Often these patients are too heterogeneous with regard to treatment, stage and standard prognostic factors to support therapeutically relevant conclusions. Many publications attempt to show that new factors are 'independently prognostic' or are more prognostic than standard factors, but these analyses often fail to identify a role of the new factors in therapeutic decision-making. Prognostic markers can be therapeutically relevant if they are developed in a focused way to identify patients whose prognosis is sufficiently good as not to require extra therapy such as with Oncotype DX [5].

Predictive classifiers identify patients who are likely or unlikely to benefit from a specific treatment. For example, HER2 amplification is a predictive classifier for benefit from trastuzumab and perhaps also from doxorubicin [6] and taxol [7]. HER2 is the target of trastuzumab, and so its use as a predictive classifier has the advantage of biological interpretability. The objective of a predictive classifier is, however, to enable one to predict accurately tumors that will or will not be responsive to a particular drug. For this purpose, biological interpretability is desirable but not essential.

Development of a predictive classifier is in some cases limited by uncertainty in the drug target, or the target that results in antitumor effect. Even in cases where the relevant target is known, it may not clear how best to measure the essentiality of the target to the pathogenesis of the tumor. For example, with trastuzumab, there was a question of whether to measure expression of the protein product or amplification of the gene. Similarly, the presence of a mutation in the kinase domain of the epidermoid growth factor receptor (*EGFR*) gene appears to be a predictive marker for response to EGFR inhibitors in patients with non-small cell lung cancer [8], although it is unclear whether EGFR amplification is a better predictive marker or whether either is sufficiently predictive for clinical use [9]. A predictive classifier may also be used to identify patients who are poor candidates for a particular drug; for example, colorectal cancer patients whose tumors have KRAS mutations may be poor candidates for treatment with EGFR inhibitors [10,11].

There are many approaches to developing predictive classifiers. They range from measuring expression of the known drug target, to whole genome expression profiling to identify a signature that distinguishes tumors that respond to the drug from those that do not. Gene expression-based classifiers generally combine the expression levels of a many genes. For example, the Oncotype DX recurrence score is a weighted average of the expression levels of 21 genes [5]. The genes to use in the classifier and the weights were determined using a training set of data to optimize predictive accuracy. For developing a predictive classifier of patients likely to benefit from a new drug, one can perform gene expression profiling of patients on Phase II trials of the drug and compare the expression profiles of responders with those of non-responders to identify the differentially expressed genes and train a completely specified predictive classifier. Dobbin and co-workers [12,13] have developed methods for planning the number of cases needed to develop effectively such a classifier. In many cases there will be too few responders in the Phase II database of the new treatment for this approach [14]. In such a case one might be able to develop a classifier that identifies patients unlikely to benefit from standard therapy because there may be many more patients available who have received standard therapy. Such a classifier could subsequently be used to focus development of the new drug on patients who are not likely to respond to standard therapy.

There is a substantial literature on the development of gene expression-based classifiers. The components of the process include identification of the genes to be included in the classifier, selecting a mathematical way of combining the expression levels of the individual genes, and training the classifier, that is, determining the weights and cut-points, on a training set of data to distinguish the responders from the non-responders [15]. The BRB-ArrayTools software provides extensive resources for development and complete cross-validation of a wide range of prognostic and predictive classifiers based on gene expression data for binary response or survival end points [16,17]. In this paper, the use of predictive classifiers in the design of prospective trials is focused on, to determine whether a new drug is effective and how its effectiveness relates to the classes defined by a predictive classifier. Challenging and multifaceted issues involved in the development of the classifier are not focused on here. In general, if a diagnostic is to be co-developed with a drug, the Phase II studies must be designed to evaluate the candidate approaches for developing a predictive classifier. Performing this evaluation, selecting one approach, developing and analytically validating the robustness and reproducibility of the classifier before launching the Phase III trials is indeed a challenge.

When the antitumor activity of the drug appears limited to a small (e.g., 33% or less) proportion of the traditionally diagnosed set of patients, however, a companion diagnostic may be essential to the effective development of the drug.

3. Enrichment design

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

With an enrichment design, a diagnostic test is used to restrict eligibility for a randomized clinical trial of a regimen containing a new drug to a control regimen [18-20]. This approach 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 [21]. Simon and Maitournam [22-24] studied the efficiency of this approach relative to the standard approach of randomizing all patients without measuring the diagnostic. They found that the efficiency of the enrichment design depended on the prevalence of test positive patients and on the effectiveness of the new treatment in test negative patients. When fewer than half of the patients are test positive and the new treatment is relatively ineffective in test negative patients, the number of randomized patients required for an enrichment design is often dramatically smaller than the number of randomized patients required for a standard design. This was the case for trastuzumab even though the immunohistochemistry assay was far from ideal and has subsequently been replaced by a FISH-based test of HER2 amplification [25]. Simon and Maitournam also compared the enrichment design with the standard design with regard to the number of screened patients. Zhao and Simon have made the methods of sample size planning for the design of enrichment trials available online [17]. The web-based programs are available for binary, survival/disease-free survival, or uncensored quantitative end points. The planning takes into account the performance characteristics of the tests and specificity of the treatment effects. The programs provide comparisons with standard non-enrichment 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 where there is such a strong biological basis for believing that test negative patients will not benefit from the new drug that including them would be unethical. In many situations, the biological basis is strong but not compelling. Our understanding of the molecular targets of the drug is sometimes flawed and there is often uncertainty about how to measure whether a target pathway is driving invasion of a specific tumor. On the other hand, we do not really want to include test negative patients in a clinical trial to show that a treatment that we do not believe will work for them actually does not work. The enrichment design does not provide data on the effectiveness of the new treatment compared with control for test negative patients. Consequently, unless there are preliminary data 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 as a medical device. Consequently, designs that include both test positive and test negative patients should be considered. The advantages and limitations of the enrichment design and other designs described here are summarized in Table 1.

4. Including both test positive and test negative patients

When a predictive classifier has been developed but there is uncertainty about its usefulness, it is generally best to include both classifier positive and classifier negative in the pivotal clinical trials comparing the new treatment with the control regimen. It is essential, however, that an analysis plan be predefined in the protocol for how the predictive classifier will be used in the analysis. It is not sufficient just to stratify the randomization with regard to the classifier without specifying a complete analysis plan. In fact, the real importance of stratifying the randomization is that it assures that only patients with adequate test results will enter the trial.

It is important to recognize that the purpose of the pivotal trial is to evaluate the new treatment 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 component. If one does any of this, then an extra Phase III trial may be needed to evaluate treatment benefit in subsets determined by the new classifier. In moving from *post hoc* correlative science to reliable predictive medicine both statisticians and clinical investigators must learn to separate strictly the data used for developing classifiers from the data used for testing treatment effects in subsets determined by those classifiers. Only by strictly honoring this principle can reliable conclusions be achieved. The process of classifier development can be exploratory but the process of evaluating treatments should not be; it should be based on testing prespecified hypotheses in prespecified patient groups [26]. In the following sections a variety of analysis strategies are described. An attempt will be made to relate these strategies to sample size planning. At this point the author cannot provide published examples of co-developments using these designs. There are few approved companion diagnostics in oncology today and those that are available were mostly developed *post hoc*. As indicated above,

Table 1. Design strategies for use of a predictive biomarker in a pivotal trial of a therapeutic.

Design	When to use	Strengths	Limitations
Enrichment	With strong biological evidence that potential treatment effectiveness is limited to test positives	Small number of randomized patients required	Does not establish utility of test Requires analytically validated test be available at start of pivotal trial
Including test negatives and positives with defined analysis plan	When enrichment design is not appropriate	Permits establishing utility of treatment and test	Requires sample size large enough to evaluate treatment in negatives and positives separately
Adaptive threshold design	When threshold for positivity of test is not established by start of pivotal trial	Permits establishing utility of treatment and test Reduces dependence on Phase II data for establishing test threshold	More complex analysis plan than with predefinition of test threshold
Adaptive signature design	When emphasis is on overall treatment effect but a fallback secondary analysis is desired	Enables test to be determined based on randomized data for patients included in pivotal trial	Limited power for testing treatment effectiveness in test positive subset Subset analysis may be based on test not previously analytically validated

prospective co-development is very challenging. There is, however, substantial current activity in co-development using structured prospective designs such as those described below.

4.1 Analysis of test negatives contingent on significance in test positives

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

With this analysis plan, the number of test positive patients required is the same as for the enrichment design, say n_E. When that number of patients is accrued, there will be ~ $n_E/prev$ total patients and ~ $n_- = (1 - prev)\ n_E/prev$ test negative patients, where *prev* denotes the proportion of test positive patients. One should make sure that the n_E is large enough that there are sufficient test negative patients for analysis. With a time-to-event end point such as survival or disease-free survival, the planning will be somewhat more complex.

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 requires ~ 88 events of test positive patients. At the time that there are E_+ events in test positive patients, there will be approximately

$$E_- = E_+\left(\frac{\lambda_-}{\lambda_+}\right)\left(\frac{1-prev}{prev}\right) \quad (1)$$

events in the test negative group. In Equation 1 the symbols λ_- and λ_+ denote the event rates in the test negative and test positive control groups at the time that there are E_+ events in the test positive group. As above, *prev* represents the proportion of test positive patients. If the test is predictive for treatment benefit but not prognostic, then the ratio of lamdas in Equation 1 will have value 1. If E_+ is 88, if the prevalence of test positive patients is 0.25 and if the test is not prognostic, then E_- will be ~ 264 at the time of analysis. This will provide ~ 90% power for detecting a 33% reduction in hazard at a two-sided significance level of 5%. 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 for the enrichment design.

4.2 Analysis determined by interaction test

The traditional approach to the 'two-way analysis of variance' problem is first to test whether there is a significant interaction between the effect of one factor (treatment versus control) and the second factor (test negative and positive). The interaction test is often performed at a threshold above the traditional 5% level. If the interaction test is not significant, then the treatment effect is evaluated overall, not within levels of the second factor. If the interaction test is significant, then treatment effect is evaluated separately within the levels of the second factor (e.g., test positive and test negative classes). This is similar to the test proposed by Sargent *et al.* [27]. In the

example described above with 88 events in test positive patients and 264 events in test negative patients, the interaction test will have ~ 93.7% power at a one-sided significance level of 0.10 for detecting an interaction with 50% reduction in hazard for test positive patients and no treatment effect in test negative patients. Computer simulations indicate that with 88 test positive patients and 264 test negative patients, the two-stage design with $\alpha_i = 0.10$ detects a significant interaction and a significant treatment effect in test positive patients in 88% of replications when the treatment reduces hazard by 50% in test positive patients and is ineffective in test negative patients.

4.3 Subset evaluated only if overall treatment effect is not significant

Simon and Wang [28] proposed an analysis plan in which the new treatment group is first compared with the control group overall. If that difference is not significant at a reduced significance level such as 0.03, then the new treatment is compared with the control group just for test positive patients. The latter comparison uses a threshold of significance of 0.02, or whatever portion of the traditional 0.05 not used by the initial test. This design was intended for situations where it was expected that the new treatment would be broadly effective, the subset analysis being a fallback option.

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, then the overall analysis will take place when there are 297 events. If the test is positive in 25% of patients and the test is not prognostic, then at the time of analysis there will be ~ 75 events among the test positive patients. If the overall test of treatment effect is not significant, then the subset test will have power 0.75 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 power 0.80 can be achieved when there are 84 events and power 0.90 can be achieved when there are 109 events in the test positive subset.

Song and Chi [29] have proposed a refinement of the significance levels used that takes into account the correlation between the test of overall treatment effect and the treatment effect within the test positive subset.

5. Adaptively modifying types of patients accrued

Wang *et al.* [30] proposed a Phase III design comparing a new treatment with a control that starts with accruing both test positive and test negative patients. An interim analysis is performed evaluating the new treatment in the test negative patients. If the observed efficacy for the control group exceeds that for the new treatment group and the difference exceeds a futility boundary, then accrual of test negative patients terminates and accrual of extra test positive patients is substituted for the un-accrued test negative patients until the originally planned total sample size is reached. Wang *et al.* show computer simulations that indicate this design has greater statistical power than non-adaptive approaches, but their design accrues 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 can be implemented, however, without the extension of trial duration resulting from substitution of test positive for test negative patients to achieve a prespecified total sample size. The trial may be planned for specified numbers of test positive and test negative patients in order to provide adequate power for separate analysis of the two subsets. For example, the tests could be powered to detect a 50% reduction in hazard in test positive patients and a 33% reduction in test negative patients. If an interim analysis indicates futility for the test negative patients, then accrual of test negative patients may cease without affecting the target sample size or analysis for the test positive patients. This approach is only likely to be useful if the time to observing a patient's end point is rapid relative to the accrual rate. For many oncology Phase III trials using survival or disease-free survival, however, most of the patients would be accrued before a meaningful futility analysis based on sufficient events could be conducted. A further limitation is the conservativeness of futility boundaries. Wang *et al.* [30] require that the futility boundary be in the region in which the observed efficacy is greater for the control group than for the new treatment group. Consequently, even if the new treatment is completely ineffective in the test negative subset, the futility analysis will be successful less than half of the time.

6. Adaptively determining threshold of test positivity

Jiang *et al.* [31] reported on a 'Biomarker Adaptive Threshold Design' for situations where a predictive index is available at the start of the trial but a cut-point for converting the index to a binary classifier is not established. With their design, tumor specimens are collected from all patients at entry but the value of the predictive index is not used as an eligibility criterion. Their analysis plan does not stipulate that the assay for measuring the index needs to be performed in real time, although such stratification could be used. Jiang *et al.* described two analysis plans. Analysis plan A begins with comparing 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 then the new treatment is considered to be effective for the eligible population as a whole. Otherwise, a second stage test is performed using significance threshold $\alpha_2 = 0.05 - \alpha_1$. The second stage test involves finding the cut-point b for the predictive index which leads to the largest treatment versus control treatment effect when restricted to patients with predictive index above b. Jiang *et al.* maximized the partial

log likelihood for proportional hazards models for survival data restricted by each candidate cut-point level in order to find b. Let $S(b)$ denote the partial log likelihood.for the treatment effect when restricted to patients with predictive index above b. Jiang *et al.* evaluated the statistical significance of $S(b)$ by randomly permuting the labels of which patients were in the new treatment group and which were controls and determining the maximized partial log likelihood for the permuted data. This is done for thousands of random permutations. If the value $S(b)$ is beyond the $1 - \alpha_2$, the percentile of this null distribution created from the random permutations, then the second stage test is considered significant. They also describe construction of a confidence interval for the optimal cut-point b using a bootstrap resampling approach.

The advantage of procedure A is its simplicity and that it explicitly separates the test of treatment effect in the broad population from the subset selection. However, the procedure takes a conservative approach in adjusting for multiplicity of combining the overall and subset tests. An alternative analysis plan B proposed by Jiang *et al.* does not use a first stage comparison of treatment groups overall. Consequently, plan B is more appropriate to settings in which there is greater expectation that treatment effect will be limited to a predictive index-defined subset. Jiang *et al.* [31] conducted a simulation study to evaluate performance of the proposed procedures. They found that procedure B was more effective than procedure A but that both were superior to the overall test ignoring the biomarker in cases where less than half of the patients benefited from the new treatment. Jiang *et al.* also provided approaches to sample size planning for the biomarker adaptive threshold designs.

7. Adaptively determining predictive biomarker

For co-development of a new drug and companion diagnostic it is best to have the candidate diagnostic completely specified and analytically validated before its use in the pivotal clinical trials. This is difficult, however, and in some cases is not feasible, particularly with multi-gene expression-based classifiers. Freidlin and Simon [32] proposed a design for a Phase III trial that can be used when no classifier is available at the start of the trial. The design provides for development of the classifier and evaluation of treatment effects in subsets determined by the classifier in a single trial. The analysis plan of the adaptive signature design is structured to preserve the principle of separating the data used for developing a classifier from the data used for evaluating treatment in subsets determined by the classifier, although both processes are part of the same clinical trial.

The analysis plan described by Freidlin and Simon is in two parts, as for the design of Simon and Wang [28] described above. 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 may have been measured 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 then a second stage of analysis takes place. The patients are divided into a training set and testing set. Freidlin and Simon used a 50 – 50 split, but other proportions can be employed. The data for patients in the training set are 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 indicated methods for identifying a subset of patients whose outcome on the new treatment is better than the control. They use machine learning methods based on screening thousands of genes for those with expression values that interact with treatment effect. When that subset is explicitly defined, the new treatment is compared with the control for the testing set patients with the characteristics defined by that subset. The comparison of new treatment with control for the subset is restricted to patients in the testing 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 $\alpha - \alpha_1$ to assure that the overall chance of a false positive conclusion does not exceed α. These thresholds can be sharpened using the methods of Song and Chi [30].

Friedlin and Simon proposed the adaptive signature design in the context of multivariate gene expression-based classifiers. The size of Phase II databases may not be sufficient to develop such classifiers before the initiation of Phase III trials [12-14]. Freidlin and Simon showed that the adaptive signature design can be effective for the development and use of gene expression classifiers if there is a very large treatment effect in a subset determined by a set of signature genes. The power of the procedure for identifying the subset is limited, however, by having to test the treatment effect at a reduced significance level in subset patients restricted to the testing set not used for classifier development.

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

The adaptive signature design provides a clear illustration of the distinction between the traditional approach of *post hoc* correlative studies and a clear separation between development of a classifier and the testing of subsets determined by the classifier. For registration trials, however, it has some clear limitations. One is the limited statistical power for the subset analysis described above. The other is the fact that the test used for the subset analysis may not have been analytically validated before its use in the clinical trial.

8. Prospective analysis of archived specimens from randomized clinical trial

The key features of a prospective clinical trial include having a carefully specified analysis plan for evaluating a focused hypothesis and ensuring that the trial is conducted in a manner that enables the hypothesis to be addressed in an unbiased manner. This involves ensuring that the right patients are accrued, that the treatment of interest and an appropriate control group are used, that randomization of treatment assignment is employed, and that an appropriate end point is measured in an unbiased manner. Retrospective studies are often exploratory exercises with no focused analysis plan using databases based on a heterogeneous mixture of patients treated outside randomized clinical trials. If, however, there is a well-defined hypothesis concerning the relationship to a patient subset defined by a predictive biomarker to the effectiveness of a treatment and if there is a previously available randomized clinical trial that might be used to test that hypothesis, then it is important to plan and conduct the analysis with the same formality and structure as one would a completely prospective clinical trial. This would involve developing a written protocol defining the analysis before accessing any data from the trial. The effectiveness of this approach depends on the availability of archived pretreatment specimens for patients in the clinical trial. The strength of evidence resulting from such a 'prospective-retrospective' study can be much greater than for the typical exploratory retrospective study. The credibility of the results will depend on a variety of factors including the proportion of patients with adequate specimens, the results of the analysis, other evidence available, and the biological rationale for the subset hypothesis.

The prospective/retrospective approach was used very effectively for evaluating whether panitumumab effectiveness in patients with metastatic colorectal cancer is limited to patients without activating KRAS mutations. A randomized clinical trial of panitumumab monotherapy versus best supportive care for patients with chemotherapy-refractory metastatic colorectal cancer was conducted before the development of the hypothesis that effectiveness might be determined by KRAS mutation [33]. The hypothesis that effectiveness of EGFR inhibitors for metastatic colorectal cancer was limited to tumors without KRAS mutations was developed in Phase II trials [10]. Mutation status for 427 of the 463 randomized patients in the trial of Amado *et al.* was evaluable using archived tissue blocks. The randomized trial was therefore reanalyzed with the focused hypothesis that KRAS mutation status effected efficacy of benefit to panitumumab. The study showed a highly significant interaction between treatment effect and KRAS mutation status on progression-free survival, a highly significant treatment effect of panitumumab on PFS for patients without KRAS mutations and no evidence of treatment effectiveness for patients with KRAS mutations. In circumstances where the treatment involved is approved and widely used, it may not be feasible to conduct a prospective randomized clinical trial; for example, doxorubicin and HER2 expression [6].

9. Expert opinion

A companion diagnostic should identify patients likely or not likely to benefit from the specific therapeutic. General disease biomarkers are not necessarily appropriate companion diagnostics. Traditional exploratory correlative science paradigms are useful for predictive biomarker discovery but do not provide an adequate framework for validated predictive medicine.

Co-development of therapeutics and diagnostics increases the complexity of all stages of the development process. Companion diagnostics are particularly important for effective development of a therapeutic in cases where one-third or fewer of the conventionally diagnosed patients are likely to benefit from the drug.

It is desirable to have a completely specified analytically validated test available at the start of the pivotal clinical trial.

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 [22,23]. Single arm Phase II data may be used to supplement biological rationale to establish the clinical validity of the test.

In cases where the biological basis for *selecting* only 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 assure that diagnostic specimens are available for all included patients. The analysis strategy for the pivotal trial should be completely specified prospectively and several viable analysis strategies are described in this paper.

Adaptively determining the threshold of positivity using the methods described by Jiang *et al.* [31] provides flexibility to the pivotal trial while preserving statistical rigor.

Adaptively determining the diagnostic test *ab initio* using a portion of the data from the pivotal trial as described by Freidlin and Simon enables the test to be based on randomized data for the kind of patients in the pivotal trial but provides limited power for identifying treatment benefit for the identified subset. This strategy can also be employed for using a portion of the pivotal trial data for refining a predefined test or for selecting among a small number of candidate tests.

Adaptively modifying the types of patient accrued to the pivotal trial based on accumulating data are best restricted for use with short-term end points [30].

Declaration of interest

The author has no conflict of interest to declare and no fee has been received for preparation of the manuscript.

Bibliography

1. Rothenberg ML, Carbone DP, Johnson DH. Improving the evaluation of new cancer treatments: challenges and opportunities. Nat Rev Cancer 2003;3:303-9
2. Paik S, Taniyama Y, Geyer CE. Anthracyclines in the treatment of HER2-negative breast cancer. J Natl Cancer Inst 2008;100:2-3
3. Pusztai L, Ayers M, Stec J, Hortobagyi GN. Clinical application of cDNA microarrays in oncology. Oncologist 2003;8:252-8
4. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979-85
5. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817-26
6. 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
7. Hayes DF, Thor AD, Dressler LG, et al. HER2 and response to paclitaxel in node-positive breast cancer. N Engl J Med 2007;357:1496-506
8. Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer 2007;7:169-81
9. Toschi L, Cappuzzo F. Understanding the new genetics of responsiveness to epidermal growth factor receptor tyrosine kinase inhibitors. Oncologist 2007;12:211-20
10. Khambata-Ford S, Garrett CR, Meropol NJ, et al. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. J Clin Oncol 2007;25:3230-7
11. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 2008;26:1626-34
12. Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. Biostatistics 2007;8:101-17
13. Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? Clin Cancer Res 2008;14:108-14
14. Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. Clin Cancer Res 2007;13:6080-6
15. Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. Br J Cancer 2003;89:1599-604
16. Simon R, Lam A, Li MC, et al. Analysis of gene expression data using BRB-ArrayTools. Cancer Informatics 2007;2:11-7
17. Biometric Research Branch (BRB). Available from: http://linus.nci.nih.gov
18. Temple RJ. Special study designs: Early escape, enrichment, studies in non-responders. Commun Stat Theory Methods 1994;23:499-531
19. Temple RJ. Enrichment designs: efficiency in development of cancer treatments. J Clin Oncol 2005;23:4838-9
20. Schilsky RL. Target practice: oncology drug development in the era of genomic medicine. Clin Trials 2007;4:163-6
21. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 2001;344(11):783-92
22. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 2005;10:6759-63
23. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and Correction. Clin Cancer Res 2006;12:3229
24. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Stat Med 2005;24:329-39
25. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 2005;353:1659-72
26. Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol 2004;15:1731-7
27. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005;23(9):2020-7
28. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. Pharmacogenomics J 2006;6:1667-173
29. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. Stat Med 2007;26:3535-49
30. Wang SJ, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat 2007;6:227-44
31. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst 2007;99:1036-43
32. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 2005;11:7872-8

33. Van-Cutsern E, Peeters M, Siena S, et al. Open-label phase III trial of panitumumab plus best supportive care compared to best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. J Clin Oncol 2007;25:1658-64

Affiliation

Richard Simon DSc
Chief
National Cancer Institute,
Biometric Research Branch,
9000 Rockville Pike,
Bethesda, MD 20892-7434, USA
Tel: +1 301 496 0975; Fax: +1 301 402 0560;
E-mail: rsimon@nih.gov