CHAPTER 11

DRUG AND PHARMACODIAGNOSTIC CO-DEVELOPMENT STATISTICAL CONSIDERATIONS

Richard Simon
Biometric Research Branch, National Cancer Institute
9000 Rockville Pike, Bethesda MD 20892-7434
E-mail: rsimon@nih.gov

To Appear in *Developing Molecular Diagnostics for Cancer*
(H Winther, JT Jorgensen eds)
Pan Stanford Publishing

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 entails many new challenges however, including development of an appropriate diagnostic, its analytical validation, clinical validation and utilization in pivotal trials that evaluate the medical utility of the new treatment. The pivotal treatment trials are of increased complexity and require careful prospective planning. Randomized clinical trials continue to be important for evaluating the effectiveness of new treatments but the utilization of the diagnostic in the design and analysis plan should be prospectively detailed. Clear separation of the data used for developing the diagnostic, including it's threshold of positivity, from the data used for evaluating treatment effectiveness in subsets determined by the diagnostic is usually essential. We review a variety of clinical trial designs for the co-development of new treatments and companion diagnostics. These include enrichment designs in which the diagnostic is used to restrict eligibility, several prospectively defined analysis plans for designs that include both test positive and test negative patients, and adaptive designs in which data from the pivotal trial is used to both refine the diagnostic and evaluate the new treatment in a manner that preserves the overall type I error level of the study.

1. Introduction

Clinical trials of new drugs have traditionally been conducted with broad patient populations in order to avoid discrepancies between the population tested and the population potentially treated with the drug. In oncology, however, this has resulted in treating many for the benefit of few. For example, only about 5% of women with estrogen receptor positive breast cancer that has not spread to the axilla require or benefit from cytotoxic chemotherapy. For prevention studies, the number treated to benefit one patient is even much more extreme. This over-treatment results in a substantial number of adverse events and expense for treatment of patients who receive no benefit. Accumulating understanding of genomic differences among tumors of the same primary site suggest that most molecularly targeted agents are likely to benefit only the patients whose tumors are driven by deregulation of the targeted pathways. Availability of improved tools for characterizing tumors biologically makes it increasingly possible to predict whether the tumor will be responsive to a particular treatment[1].It is important that new drugs be developed 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 from a 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 we will address some of the issues in the design of prospective phase III clinical trials for new treatments and companion diagnostic tests.

2. Predictive Biomarkers, Prognostic Biomarkers and Surrogate Endpoints

A *biomarker* is any measurement made on a biological system. Biomarkers are used for very different purposes, and this often leads to confusion in discussions of biomarker development, use and validation. In its most common usage a biomarker is a measurement that tracks disease pace; increasing as disease progresses, holding constant as a disease stabilizes and decreasing as disease regresses. There are many uses for such endpoint biomarkers in developmental studies for establishing proof of concept, dose selection, and identifying the patients most suitable for inclusion in pivotal trials. In some cases that us also interest in using and endpoint biomarker in phase III trials as a surrogate for clinical outcome. The standards for validation of a surrogate endpoint are stringent; however. It is not sufficient to demonstrate that the biomarker value is correlated with clinical outcome. It is necessary to show that treatment that impacts the biomarker value also impacts clinical outcome. This requires analysis of a series of randomized clinical trials, showing that the differences in biomarker change between the randomized treatment group is concordant with the differences in clinical outcome [2-4]. These standards are stringent because of the key role of the phase III trial endpoint in claims of effectiveness. There are well known examples where biomarkers of disease pace were not valid surrogate endpoints pf clinical outcome. Because of the stringency of the requirements for establishing a biomarker as a valid surrogate endpoint, it is often best to perform phase III trials totals using standard measures of clinical outcomes as endpoint.

Predictive biomarkers are pre-treatment measurements used to characterize the patient's disease in order to determine whether the patient is a good candidate for receiving a particular therapy. The term predictive denotes predicting outcome to a specific treatment. This is in contrast to *prognostic biomarkers* which are correlated with outcome of untreated patients or with the survival of a heterogeneously treated group of patients. 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*.[5] identified 939 publications over a 20 year period on prognostic factors in breast cancer but only four factors (ER, PR, HER2, and Oncotype DX) are recommended for use by the American Society of Clinical Oncology. Prognostic factors are often not used unless they help with therapeutic decision making. Most prognostic factor studies are conducted using a convenience sample of patients whose tissues are available[6]. The studies are often not focused on a particular medical decision facing physicians and hence the resulting prognostic actors identified have little therapeutic relevance. 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. The greatest advantage of using tissue specimens derived from patients in a clinical trial is that it tends to restrict the stuffy to a medical context from which therapeutically relevant biomarkers can be developed. The fact that patients in clinical trials are uniformly staged and adequately followed is an important bonus. Prognostic biomarkers can be therapeutically relevant if they are developed based on specimens selected to identify uniformly staged patients who do not need additional therapy following a standard regimen.

Predictive biomarkers identify patients whose are likely or unlikely to benefit from a specific treatment. For example, HER2 amplification is a predictive biomarker for benefit from herceptin and perhaps also from doxorubicin[7] and taxol[8]. The presence of a mutation in the kinase domain of the epidermoid growth factor receptor (EGFR) gene may be a predictive marker for response to EGFR inhibitors[9], although it is unclear today whether EGFR amplification is a better predictive marker or whether either is sufficiently predictive for clinical use[10]. A predictive biomarker may 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.

Some predictive biomarkers are 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 to best measure the essentiality of the target to the pathogenesis of a specific tumor. For example, with trastuzumab, there was a question of whether to measure expression of the protein product or amplification of the gene. In other cases the options will be more numerous. If a diagnostic is to be co-developed with a drug, the phase II studies must be designed to evaluate the candidate assays available, to select one, and then to perform analytical validation of the robustness and reproducibility of the assay prior to launching the phase III trial.

3. Development and Analytical Validation of Predictive Biomarker Classifiers

In this chapter we will focus on the use of predictive biomarker classifiers in the design of pivotal clinical trials. The term *classifier* indicates that the biomarker can be used to classify patients. We will generally be interested in classifying patients as either good candidates for the new drug or not good candidates, i.e. binary classifiers. If we were advising patients about their likelihood of benefit from a treatment, and probability of benefit or an index might be more informative than a binary classifier. The development of such a predictor would, however, require much more extensive data than generally available prior to performing the pivotal trial (s). We shall restrict ourselves here to binary classifiers that can be used to select patients for inclusion or exclusion from the pivotal trials.

Predictive binary classifiers can be of many types. The simplest might reflect presence or absence of a point mutation in the EGFR gene, amplification of the HER2 gene, or over-expression of the protein product of a gene. At the other extreme, the binary classifier may be based on the expression levels of a large number of genes. In such cases the component genes are generally selected for their correlation with response or patient outcome. The component genes do not themselves constitute the classifier; they must be combined in some completely defined manner. In many cases the drug being developed has multiple molecular targets and it is not known for certain which targets are the most important in treating tumors of a given primary site. Even in cases where the therapeutic target is known with greater confidence, there may be uncertainty about how best to measure the target; e.g. based on protein expression, transcript expression, gene mutation or gene amplification. There is an urgency during the early phases of clinical development to develop a diagnostic test will be used for enhancing the pivotal clinical trials of the new drug. The test should be selected from among the candidate tests to best identify those tumors which are likely to respond to the new drug. Puzstai *et al.*[11] have indicated that a small phase 2 database may not be sufficient to develop a classifier predictive of response based on de-novo whole genome expression profiling and suggested a strategy of development based on candidate genes. Dobbin *et al.* indicated that for whole genome expression profiling, a training set consisting of at least 20-30 responders and at least that many non-responders is desirable[12]. With molecularly targeted drugs, candidate genes will often be known. When adequate clinical data is not available, expression profiles of human tumor cell lines responsive to a new drug may

provide candidate genes[13]. Even with candidate genes on which to develop a classifier for predicting response to the new drug, however, a larger phase 2 database than has traditionally been available will generally be needed. Co-development of a new drug and a predictive diagnostic is a more complex endeavor than traditional development of a drug for use in a broad patient populations and it may require increased resources. One hopes that these increased resources will provide a greater chance of successful development of a drug and diagnostic.

In addition to developing a diagnostic classifier that predicts those patients likely to benefit from the new drug, the pre-pivotal developmental phase should generally establish a threshold of positivity and *analytical validation* of the test. With many such diagnostics there is no gold standard and hence analytical validation should mean that the test is reproducible and robust. That is, the test is robust to sample handling and laboratory variation. If the test is repeated with different samples of the same tumor or repeated on different days in the same or different laboratories, then the resulting classification should be unchanged.

3.1 Multivariate Gene Expression Classifiers

Two kinds of gene expression based classifiers are frequently used. Both require a training set of data consisting of pre-treatment expression levels for patients treated with the drug. The signature genes that are differentially expressed between the responders and non-responders are identified. The first type of classifier is based on a weighted average of expression for the most differentially expressed genes. Many of the commonly used classifier types are based on such weighted averages. These include Golub's weighted voting classifier [14], the combined covariate predictor [15], Fishers linear discrimant and diagonal linear discrimant analysis [16], support vector markers with inner product kernel [17], naive Bayes classifier [18], and perceptrons [19]. The methods differ in how they define the weights. Using the training date to define the eights and threshold results in a completely specified binary classifier.

The second kind of binary classifier widely used for gene expression data are the non-parametric distance based methods. These include nearest neighbor, k-nearest neighbor, nearest controid and shrunken centroid classifiers [16, 20]. These methods also use signature genes that are differentially expressed between responders and non-responders. A distance metric is adopted for measuring the similarity or dissimilarity between expression profiles with regard to the signature genes. Usually Euclidean distance or correlation is used. If a new patient is to be classified based on a training set of expression profiles of patients who were previously treated, one finds the training sample to which that the new patient profile is most similar. That training sample is called the "nearest neighbor" of the profile of the new patient. If that nearest neighbor was a responder, then the new patient is predicted to be a responder; if the nearest neighbor was a non-responder, then the new patient is predicted to be a non-responder. The k-nearest neighbor algorithm is similar to except a majority vote of the classes of the k closest profiles to that of a new patient are used for prediction. Nearest centroid and shrunken centroid methods work similarly.

The predictive accuracy of the binary classifier must be evaluated on a separate set of data. Using the same data to develop a classifier and evaluate its accuracy results in very misleading results unless special methods of complete cross-validation methods are used [21]. Unfortunately, cross validation methods are used improperly in many cases [22].
One should also establish that the predictive accuracy of the classifier is better than can be achieved without the gene expression data. The BRB-Array Tools Software [23] provides a convenient integrated environment for identifying signature genes, developing weighted average and non-parametric distance based classifiers, validly evaluating prediction accuracy, sensitivity

and specificity and testing whether the predictive accuracy is statistically significant . The software is available at http://linus.nci.nih./gov/brb. Additional details about the development of predictive biomarker classifiers based on gene expression data are available from [16, 24, 25].

4. Clinical Validation of Predictive Biomarkers

If a predictive biomarker is developed to identify patients who are likely to respond to a specified treatment, then clinical validation usually involves establishing that the completely specified test actually does predict response for a set of patients independent of the patients used for the development of the test. This objective does not generally require a randomized clinical trial. A sufficiently large single arm trial of patients receiving the new drug may be sufficient to establish clinical validity of the test for predicting response, but may very well not be sufficient to establish either effectiveness of the new treatment or medical utility of the test. For those objectives, a phase IIIl clinical trial is generally required in which patients are randomized to receive the new treatment or a control treatment, with all patients receiving the new test.

5. Use of Predictive Biomarkers in the Design of Phase III Clinical Trials

The objective of a pivotal phase III clinical trial is to evaluate whether a new drug, given in a defined manner, has a medical utility for a prospectively specified patient population. The role of a predictive biomarker classifier is to refine the population of patients. The process of biomarker classifier development may be exploratory and subjective, but the use of the classifier in the pivotal trial must not be. If the data from a phase III trial is to be used to develop or refine a biomarker classifier, then treatment hypotheses involving that classifier should be tested in a separate trial. One exception is the adaptive trial design of Friedlin and Simon [26] where some data from a phase III trial is used to develop a classifier and that data is excluded from the data for that same trial that is used to test a treatment hypothesis in the subset of patients defined as positive by that classifier.

6. Enrichment Design

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. 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. Simon and Maitournam[27-29] 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 has subsequently been replaced by a FISH based test of HER2 amplification. Simon and Maitournam also compared the enrichment design to 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 on line at http://linus.nci.nih.gov/brb/ . The web based programs are available for binary, survival/disease-free survival, or uncensored quantitative endpoints. The planning takes into account the performance characteristics of the tests and specificity of the treatment effects. The programs provide comparisons to 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 than including them in would raise ethical concerns. In many situations, the biological basis is strong but not compelling. The enrichment design does not provide data on the effectiveness of the new treatment compared to control for test negative patients. Consequently, unless there is preliminary 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.

7. Including Both Test Positive and Test Negative Patients

When test positive and test negative patients are included in the randomized clinical trial comparing the new treatment to a control, it is essential that an analysis plan be pre-defined in the protocol for how the predictive classifier will be used in the analysis. It is not sufficient to just stratify the randomization with regard to the classifier without specifying a complete analysis plan. In fact, the main 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 this design is to evaluate the new treatment in the subsets determined by the pre-specified 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 additional 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 strictly separate the data used for developing classifiers from the data used for testing treatment effects in subsets determined by those classifiers. The process of classifier development can be exploratory, but the process of evaluating treatments should not be; it should be based on testing pre-specified hypotheses in pre-specified patient groups. In the following sections we will describe several analysis strategies and relate these strategies to sample size planning.

7.1 Analysis of Test Negatives Contingent on Significance in Test Positives

The simplest analysis plan consists of separate comparisons of the new treatment to 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 statistical significance test in negative patients. If, however, the treatment is significantly better than control in test positive patients, then compare treatment to 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 are accrued, there will be approximately n_E/prev total patients and approximately $n_- = (1\text{-prev})\ n_E$/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 a sufficient number of test negative patients for analysis. With a time-to-event endpoint like 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 about 88 events of test positive patients. At the time that there are E_+ events in test positive patients, there will be approximately

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

events in the test negative group. In expression (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. If the test is predictive for treatment benefit but not prognostic, then the ratio of lamda's in (1) will have value 1. If E_+ is 88, if the prevalence of test positive patients is .25 and if the test is not prognostic, then E_- will be approximately 264 at the time of analysis. 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 to 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.

7.2 Analysis Determined by Interaction Test

The traditional approach to the "two-way analysis of variance" is to first 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 versus control comparison is evaluated overall, not within levels of the second factor. If the interaction test is significant, then treatment versus control comparison 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[30]. In the example described above with 88 events in test positive patients and 264 events in test negative patients, the interaction test will have approximately 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 α_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.

7.3 Test Positive Subset Evaluated Only if Overall Treatment Effect is not Significant

Simon and Wang [31] proposed an analysis plan in which the new treatment group is first compared to 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 to 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 .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 approximately 75 events among the test positive patients. If the overall test of treatment effect is not significant, then the subset test will have 75% power for detecting a 50% reduction in hazard at a two-sided .02 significance level. By delaying the treatment evaluation in the test positive patients 80% power can be achieved when there are 84 events and 90% power can be achieved when there are 109 events in the test positive subset. .

Song and Chi [32] 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.

8. Adaptively Deterimining Threshold of Test Positivity

Jiang et al. [33] 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 criteria. Their analysis plan does not stipulate that the assay for measuring the index needs to be performed in real time, though regulators may prefer that the index be used to stratify the randomization between the new treatment and control. Jiang et al. described two analysis plans. Analysis plan A begins with comparing outcomes for all patients receiving the new treatment to those for all control patients. If this difference in outcomes is significant at a pre-specified significance level α_1 then the new treatment is considered effective for the eligible population as a whole.

Otherwise, a second stage test is performed using significance threshold $\alpha_2 = .05 - \alpha_1$. 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 $1-\alpha_2$ 'th 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 re-sampling 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.[33] 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.

9. 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 prior to its use in the pivotal clinical trials. This is difficult, however, and in some cases may not be feasible, particularly with multi-gene expression based classifiers. Freidlin and Simon[26]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 preserves the principal 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[31] described above. At the conclusion of the trial the new treatment is compared to the control overall using a threshold of significance of α_1 which is somewhat less than the traditional α=.05. 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 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 is used to define a single subset of patients who are expected to be most likely to benefit from the new treatment compared to 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 to the control for the testing set patients with the characteristics defined by that subset. The comparison of new treatment to 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 to control for the subset uses a threshold of significance of $\alpha - \alpha_1$ in order 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[32].

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[11, 12, 34]. 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 very stringent 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. 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 over-expression 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.

10. Conclusions

Developments in cancer genomics and biotechnology are dramatically changing the opportunities for development of more effective cancer therapeutics and molecular diagnostics to guide the use of those drugs. These opportunities can have enormous benefits for patients and for containing health care costs. One of the greatest opportunities is developing predictive biomarkers of the patients who require treatment and are likely to benefit from specific drugs. Co-development of drugs and companion diagnostics adds complexity to the development process however. 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. 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. This paper has attempted to begin to touch on some of these approaches.

REFERENCES

[1] Paik S, Taniyama Y, Geyer CE. Anthracyclines in the treatment of HER2-negative breast cancer. Journal of the National Cancer Institute. 2008;100:2-3.

[2] Torri V, Simon R, Russek-Cohen E, Midthune D, Freidman M. Relationship of response and survival in advanced ovarian cancer patients treated with chemotherapy. Journal of the National Cancer Institute. 1992;84:407.

[3] Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate endpoints in multiple randomized clinical trials with failure time end points. Journal of the royal statistical Society, Series C, Applied Statistics. 2001;50(4):405.

[4] Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. Statistics in Medicine. 2004;24:163-82.

[5] Pusztai L, Ayers M, Stec J, Hortobagyi GN. Clinical application of cDNA microarrays in oncology. The Oncologist. 2003;8:252-8.

[6] Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. British Journal of Cancer. 1994;69:979-85.

[7] Gennari A, Sormani MP, Pronzato P, Puntoni M, Colozza M, Pfeffer U, et al. HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: A pooled analysis of randomized trials. Journal of the National Cancer Institute. 2008;100:14-20.

[8] Hayes DF, Thor AD, Dressler LG, al. e. HER2 and response to paclitaxel in node-positive breast cancer. New England Journal of Medicine. 2007;357:1496-506.

[9] Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nature Reviews Cancer. 2007;7:169-81.

[10] Toschi L, Cappuzzo F. Understanding the new genetics of responsiveness to epidermal growth factor receptor tyrosine kinase inhibitors. The Oncologist. 2007;12:211-20.

[11] Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. Clinical Cancer Research. 2007;13:6080-6.

[12] Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? Clinical Cancer Research. 2008;(in press).

[13] Potti A, Dressman HK, Bild A, al e. Genomic signatures to guide the use of chemotheraputics. Nature Medicine. 2006;12:1294-300.
[14] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression modeling. Science. 1999;286:531-7.
[15] Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biololgy. 2002;9:505-12.
[16] Simon R, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and Analysis of DNA Microarray Investigations. New York: Springer Verlag 2003.
[17] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Science. 2001;98:15149-54.
[18] Hand DJ, Yu K. Idiot's Bayes-Not so stupid after all? International Statistical Review. 2001;69(3):385-98.
[19] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001;7:673-9.
[20] Tibshirani R, Hastie T, al. e. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Science. 2002;99:6567-72.
[21] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute. 2003;95:14-8.
[22] Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. Journal of the National Cancer Institute. 2007;99:147-57.
[23] Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-ArrayTools. Cancer Informatics. 2007;2:11-7.
[24] Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. British Journal of Cancer. 2003;89:1599-604.
[25] Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology. 2005;23:7332-41.
[26] Freidlin B, Simon R. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clinical Cancer Research. 2005;11:7872-8.
[27] Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research. 2005;10:6759-63.
[28] Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: Supplement and Correction. Clinical Cancer Research. 2006;12:3229.
[29] Maitournam A, Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine. 2005;24:329-39.
[30] Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of Clinical Oncology. 2005;23(9):2020-7.
[31] Simon R, Wang SJ. Use of genomic signatures in therapeutics development. The Pharmacogenomcs Journal. 2006;6:1667-173.
[32] Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. Statistics in Medicine. 2007;26:3535-49.
[33] Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. Journal of the National Cancer Institute. 2007;99:1036-43.
[34] Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. Biostatistics. 2007;8:101-17.