Guidelines for the Design of Clinical Studies for Development and Validation of Therapeutically Relevant Biomarkers and Biomarker Based Classification Systems

Richard M. Simon

To Appear in *Biomarkers in Breast Cancer*

D Hayes & G. Gasparini editors

Running Title: Designing clinical studies for biomarker development

Key Words: Biomarkers, Microarrays, Classification systems, Clinical trial design

Richard M. Simon, D.Sc.
Chief, Biometric Research Branch
National Cancer Institute
6130 Executive Blvd. Room 8134
Rockville MD 20852
rsimon@nih.gov
tel (301)496-0975
fax (301)402-0560

Abstract

Standards for the development of therapeutically relevant biomarkers and biomarker based classification systems are lacking. The literature of prognostic marker studies for breast cancer is inconsistent and few such markers have been adopted for widespread use in clinical practice. This is problematic as many patients are overtreated and many others are treated ineffectively. The deficiencies in clinical development of biomarkers may become more severe as DNA microarrays and proteomic technologies provide many new candidate markers and therapeutics become more molecularly targeted. In this chapter we address some common problems with developmental marker studies and provide recommendations for the design of clinical studies for the development and validation of robust, reproducible and therapeutically relevant biomarkers and biomarker based classification systems. The design of validation studies is addressed for (i) identifying node negative breast cancer patients who do not require systemic chemotherapy; (ii) identifying node positive breast cancer patients who do not benefit from standard chemotherapy; and (iii) identifying node positive breast cancer patients who benefit from a new molecularly targeted therapeutic.

1. Introduction

Breast cancer is a heterogeneous set of diseases. Although substantial progress has been made in the treatment of breast cancer, many patients are over-treated and many undergo intensive chemotherapy with little apparent benefit. The literature on prognostic factors in breast cancer, although voluminous, is inconsistent [1]. The process of how to develop biomarkers that are robust, reproducibly measured, and therapeutically has not been well established. Although many prognostic factors have been studied, treatment selection has remained based primarily on the traditional components of TNM stage and hormone receptor levels. This discrepancy between an inconsistent research literature and clinical practice will become even more problematic as DNA microarray and proteomic technologies provide new markers and therapeutics become more molecularly targeted. The objectives of this chapter are to provide information that facilitates the development of biomarkers for selection of the best treatment for each patient. We will use the term *biomarker* to include predictive classification systems based on protein or RNA transcript profiles measured using technology such as DNA microarrays.

2. Pitfalls in Developmental Studies

Most biomarkers are developed using archived tumor specimens, and many of the problems that exist in the marker literature derive from the retrospective nature of these studies. Clinical drug trials are generally prospective, with patient selection criteria, primary endpoint, hypotheses and analysis plan specified in advance in a written protocol. The consumers of clinical trial reports have been educated to be skeptical of *data dredging* to find something “statistically significant” to report in clinical trials. They are skeptical of analyses with multiple endpoints or multiple subsets, knowing that the chances of erroneous conclusions increase rapidly once one leaves the context of a focused single hypothesis clinical trial. Marker studies are generally performed with no written protocol, no eligibility criteria, no primary endpoint or hypotheses and no defined analysis plan. The patient population is often very heterogeneous and represents individuals for whom archived specimens are available. The patients are often not treated in a single clinical trial and represent a mixture of stages. Consequently, the overall population often does not represent a therapeutically meaningful group and the biomarkers identified may be of prognostic relevance, but less likely to be of predictive relevance for selecting therapy. Often the marker may be prognostic because it is correlated with disease stage or some other known prognostic marker. Broad populations are also often heterogeneously treated and so finding that a marker is prognostic in such a population may be difficult to interpret. Prognostic markers that do not have therapeutic implications are rarely used. The heterogeneous nature of the population also often results in multiple subset analyses of more therapeutically meaningful sub-populations. With multiple analyses, the chance of false positive conclusions increases. Many biomarker studies perform analyses for many candidate biomarkers and several endpoints as well as for various patient subsets. Consequently, the chance for erroneous conclusions increases multiplicatively. The multiplicity problem is even more severe when one considers that there are usually multiple ways of quantifying biomarker level and many possible mathematical models for combining biomarker measurements.

Many of the problems that have hindered the development and acceptance of predictive single protein biomarkers also apply for biomarkers based on DNA microarray expression profiles[2]. There are multiple platforms and protocols for measuring expression profiles, and microarray research studies almost never evaluate inter-laboratory assay reproducibility. Microarray expression profiles in research studies are generally performed at one time so that reagent variability is minimized and it is almost never demonstrated that the models are predictive for tumor specimens collected and assayed at other times. This is of particular concern for printed cDNA microarrays where there may be substantial variability among batches of printed slides and batches of reference RNA.

Because of the number of genes available for analysis, microarray data can be a veritable fountain of false findings unless appropriate statistical methods are utilized. For example, in comparing expression profiles of 10,000 genes for tumor specimens selected from patients who have responded to a specified treatment to those for non-responders, the expected number of false positive genes that are statistically significantly ($p<0.05$) differentially expressed between the two groups is 500. This is true regardless of whether the expression levels for different genes are correlated. Consequently, more stringent methods for assessing differential expression must be used. Some studies do not use statistical significance at all and just identify genes as differentially expressed based on fold-change statistics; that is the ratio of the average expression level in responders to the average in non-responders, ignoring variability entirely. Others base their analyses on visual inspection of graphical data displays. Such methods are clearly problematic.

The unstructured nature of retrospective studies of biomarkers would not be so problematic if they were followed by structured prospective validation studies that tested specific hypotheses about predictive biomarkers. Such prospective trials are rarely performed, however, because they are difficult to accomplish. Consequently, before discussing the design of such prospective trials, we will make some suggestions about a more structured approach to retrospective studies.

3. Structured Retrospective Studies

There is a role for exploratory studies in which multiple biomarkers and multiple ways of combining biomarkers into predictive models are examined so long as one has an adequate way of evaluating the result. A major problem with many retrospective studies is that they attempt to use the same set of data to both develop hypotheses (biomarkers) and to test those hypotheses. This problem is particularly severe when the number of candidate hypotheses examined in the exploratory stage is large.

In trying to determine which genes are differentially expressed in comparing responders to a given therapy to non-responders, the number of hypotheses equals the number of genes examined. The Bonferonni method of adjusting for multiple testing requires that the p value calculated for comparing expression of a specific gene i in responders to non-responders, say p_i, be adjusted based on the number of genes (N) examined. For microarray studies, N could be 10,000 or greater. The Bonferonni method tries to eliminate all false positives. For microarray studies, less conservative methods that control the number of *false discoveries* (false positives), or the proportion of claimed positives that are false positives (*false discovery rate*) [3]. These same ideas apply if, for example, we are examining which genes are prognostic for survival or disease-free survival on a particular treatment.

For assessing statistical significance, adjustments such as those described in the preceding paragraph can be applied to adjust for the fact that we don't have a specific hypothesis to test, but rather are in a hypothesis development mode. The adjustment is based on treating the problem as one of testing all possible hypotheses. For retrospective biomarker studies in which a number of biomarkers are examined, such adjustments to statistical significance should be applied. In many cases, however, statistical significance is not the best measure of biomarker value. A better measure is the extent to which the biomarker model enables us to predict whether the patient will respond to the treatment [4].

For binary outcomes like response and non-response, the best measure of predictive accuracy is the number of correct predictions. For quantitative outcomes such as survival or disease-free survival, measurement of predictive accuracy is more complex. In many cases, it is reasonable to approximate quantitative outcomes in a binary manner, good outcome or poor outcome. In other cases, measures such as described by Korn and Simon [5] are used.

It is not valid to use the same set of data for selecting a predictive marker or developing a predictive model and for measuring predictive accuracy. The estimate of predictive accuracy computed on the same data used to select the marker or develop the model is called the *resubstitution* estimate and is known to be biased [6]. The bias is extreme when the number of candidate markers is larger than the number of cases. For example, Simon et al. [6] showed that for two classes (e.g. responders and non-responders) that have no genes that are truly differentially expressed in microarray expression profiles of thousands of genes, one can almost always find a predictive model that has a resubstitution estimate of accuracy of 100%. Such a model would be useful for future data, but would appear to give perfect predictions for the cases used to develop the model.

How can we develop a proper estimate of the accuracy of class prediction for future samples? For a future sample, we will apply a fully specified predictor developed using the data available today. If we are to emulate the future predictive setting in developing our estimate of predictive accuracy, we must set aside some of our samples and make them completely inaccessible until we have a fully specified predictor that has been developed from scratch without utilizing those set aside samples.

To properly estimate the accuracy of a predictor for future samples, the current set of samples must be partitioned into a training set and a separate test set. The test set emulates the set of future samples for which class labels are to be predicted. Consequently the test samples cannot be used in any way for the development of the prediction model. This means that the test samples cannot be used for estimating the parameters of the model and they cannot be used for selecting the gene set to be used in the model. It is this later point which is often overlooked.

The most straightforward method of estimating the accuracy of future prediction is the *split-sample* method of partitioning the set of samples into a training set and a test set as described in the previous paragraph. Rosenwald et al. [7] used this approach successfully in their international study of prognostic prediction for large cell lymphoma. They used two thirds of their samples as a training set. Multiple kinds of predictors were studied on the training set. When the collaborators of that study agreed on a fully specified prediction model, they accessed the test set for the first time. On the test set there was no adjustment of the model, re-defining of cutoff values or fitting of parameters. They merely used the samples in the test set to evaluate the predictions of the model that was completely specified using only the training data.

Cross-validation is an alternative to the split sample method of estimating prediction accuracy [8]. Cross-validation can only be used when there is a well defined algorithm for predictive model development. In such cases, cross-validation can be more efficient than the split-sample method for estimating prediction accuracy. There are several forms of cross-validation. Here we will describe *leave-one-out cross-validation (LOOCV)* in the context of a class predictor based on gene expression levels determined by DNA microarray analysis. LOOCV starts like split-sample cross validation in forming a training set of samples and a test set. With LOOCV, however, the test set consists of only a single sample; the rest of the samples are placed in the training set. The sample in the test set is placed aside and not utilized at all in the development of the class prediction model. Using only the training set, the informative genes are selected and the parameters of the model are fit to the data. Let us call M_1 the model developed with sample 1 in the test set. When this model is fully developed, it is used to predict the class of sample 1. This prediction is made using the expression profile of sample 1, but obviously without using knowledge of the true class of sample 1. Symbolically, if $\underline{x}_1$ denotes the complete expression profile of sample 1, then we apply model M_1 to $\underline{x}_1$ to obtain a predicted class $\hat{c}_1$. This predicted class is compared to the true class label c_1 of sample 1. If they disagree, then the prediction is in error. Then a new training set – test set partition is created. This time sample 2 is placed in the test set and all of the other samples, including sample 1, are placed in the training set. A new model is constructed from scratch using the samples in the new training set. Call this model M_2 . Model M_2 will generally not contain the same genes as model M_1. Although the same algorithm for gene selection and parameter estimation is used, since model M_2 is constructed from scratch on the new training set, it will in general not contain exactly the same gene set as M_1. After creating M_2, it is applied to the expression profile $\underline{x}_2$ of the sample in the new test set to obtain a predicted class $\hat{c}_2$. If this predicted class does not agree with the true class label c_2 of the second sample, then the prediction is in error.

The process described in the previous paragraph is repeated *n* times where *n* is the number of biologically independent samples. Each time it is applied, a different sample is used to form the single-sample test set. During the steps, *n* different models are created and each one is used to predict the class of the omitted sample. The number of prediction errors is totaled and reported as the leave-one-out cross-validated estimate of the prediction error.

At the end of the LOOCV procedure you have constructed n different models. They were only constructed in order to estimate the prediction error associated with the type of model constructed. The model that would be used for future predictions is one constructed using all n samples. That is the best model for future prediction and the one that should be reported in the publication. The cross-validated error rate is an estimate of the error rate to be expected in use of this model for future samples assuming that the relationship between class and expression profile is the same for future samples as for the currently available samples. With two classes, one can use a similar approach to obtain cross-validated estimates of the sensitivity, specificity.

Leave-one-out cross validation is applicable only in settings where there is an algorithm for the development of a predictive model. In many studies, the analysis is less algorithmic and many kinds of prediction models are explored. For such studies, it is best to use the split sample approach of setting aside at least one third of the samples as a test or validation set. The samples in the test set should not be used for any purpose other than testing the final model developed in the training set. Specifically, the test set samples should not be used for limiting the set of genes to be considered in detail in the training set. The samples in the test set should not be accessed until a single model is identified based on training set analyses as *the* model to be tested.

LOOCV can be used to evaluate risk group predictors using survival or disease-free survival data. Suppose we wish to identify patients in a low-risk group with 10 year disease-free survival greater than 90%. Consider the leave-one-out training set in which observation i is left out in the test set. A disease-free survival model M_i is developed for the training set. For example, the model might be a proportional hazards regression model which predicts disease-free survival based on the expression profile and/or standard prognostic factors. The model M_i is applied to the left-out specimen i to obtain a prediction of the probability that the i'th patient has 10-year disease-free survival greater than 90%. Let $y_i = 1$ if this probability is greater than 50%. This process is repeated for all of the leave-one-out training sets. Then, the Kaplan-Meier disease-free survival curve estimate is computed and plotted for the patients predicted to be of very low-risk, those with $y_i = 1$. The adequacy of the model is judged by whether the estimated 10-year disease-free survival for the identified low-risk group is in fact in excess of 90%. An approach similar to this was used for developing a classification system based on survival for patients with renal cancer by Vasselli et al. [9]

One of the common errors in retrospective studies of biomarkers is that the statistical significance of the biomarker is evaluated rather than the predictive accuracy of the biomarker [4]. We have indicated above how predictive accuracy can be evaluated in a manner that avoids the bias of the re-substitution estimate. But even this is not sufficient. New biomarkers are often correlated with existing prognostic factors. The retrospective study must provide strong evidence that the new marker is substantially more predictive than the currently available prognostic factors. This can be addressed by computing the split-sample or cross-validated error rate for a model consisting of current prognostic factors and then computing the split-sample or cross-validated error rate for a model consisting of current prognostic factors plus the new candidate markers. Only if the latter is substantially greater than the former with regard to a therapeutically relevant prediction will a prospective validation study be warranted.

4. Validation Studies

Assuming that the initial study is performed properly with attention to the statistical principles described in previous section, it might be considered a phase II study, and the next step should be to conduct a phase III study that is focused on testing the specific classifier developed by the initial study [10]. The phase III study should be conducted with a written protocol. The phase III trial should be designed to test the biomarker classifier developed in the previous study. The classifier should be fully specified in the protocol. If the biomarker is expression profile based, the specification must include the genes used, the mathematical form of the classifier, parameter values and cut-off thresholds for distinguishing the classes or prognostic groups.

The phase III study should attempt to perform the assays in a manner as similar as possible to the way it would be performed broadly outside of a research setting if the diagnostic classifier were adopted. Consequently, attention is required in determining whether the same platform should be used for the phase III trial as for the phase II trial. If the platform is changed, then clearly some intermediate study will be needed to translate the classification algorithm from use on the phase II platform to the platform used in the phase III trial.

Even if there is not a change in platform, an intermediate study may be required to prepare the classifier for use with multiple laboratories performing the assay. The phase II trial may have had all of the assays performed at a single location by a research laboratory and it may be advisable to conduct the phase III trial in a manner more similar to the way it would be performed if the classifier were adopted for national use. Generally this will mean that several laboratories will be conducting the assays. Consequently, the protocol for the phase III study should specify procedures to be used for conducting the assay. It is also useful to conduct intermediate studies of inter-laboratory reproducibility of the assays. Unless inter-laboratory reproducibility is sufficiently high, it is not advisable to proceed with the phase III trial.

If the biomarker classifier was developed using a dual-label microarray platform, then use of the classifier in other laboratories requires that they use the same common reference RNA as was used for the initial study. Since different batches of the common reference will be utilized for classifying subsequent patients, calibration studies will generally be required to ensure that the expression profile of the common reference does not change and to adjust the classifier for small changes.

Conducting the validation study as a prospective trial is desirable for many reasons. One can never be sure that the patients for whom one has adequate preserved tissue are representative of the population of patients presenting for treatment. It is difficult to assure that a retrospective cohort was adequately staged and treated, and the data available may be incomplete. It is also difficult to assess whether a diagnostic procedure is practical unless it is studied in the real-time context of presenting patients who need to be evaluated and treated. Prospective accrual is also important for evaluating the diagnostic classifier in the context of real-time tissue handling. Table 1 from Simon and Altman [10] indicates some important design features of prospective validation studies.

The objective of the validation trial of a predictive marker is to test the hypothesis that the marker is useful for treatment selection. This is often a more complex objective than validation of a prognostic marker, where the objective is to determine whether the marker can separate the uniformly staged and treated patients into groups of differing outcome. There are some cases, discussed below, where prognostic markers are also predictive markers. Our focus is on predictive markers and we will consider three breast cancer scenarios.

4.1 Identifying Node Negative Patients Who Do Not Require Chemotherapy

Our first scenario is a putative marker for identifying node negative patients whose prognosis on local therapy and possibly Tamoxifen is so good that they do not require chemotherapy. The retrospective study for development of such a marker would have probably been based on archived tumors of node negative patients who did not receive chemotherapy. A tissue microarray of a large number of such specimens, with associated clinical follow-up data, can provide a valuable resource for ensuring that the marker is sufficiently promising to warrant evaluation in a prospective clinical trial if the classifier is not RNA transcript profile based. A marker of this type meets the definition of a prognostic marker, but it can also be a predictive marker if it enables us to determine which node negative patients do not require chemotherapy.

The theoretically optimal trial design would be to randomize candidate node negative patients to receive or not receive chemotherapy and then to validate whether the marker identifies those who do not benefit from chemotherapy. The candidate node negative patients might be those with tumors 1-3 cm in diameter without known poor prognostic features such as hormone receptor negativity. This is probably not a feasible approach, however, because chemotherapy has already been established as being effective for much of the candidate population.

An alternative study design is to withhold chemotherapy from a subset of node negative patients selected based on marker status to be of particularly low risk. If their outcomes were sufficiently good relative to some standard, then the marker would be accepted as useful. The standard might be based on outcomes for node negative patients that are similar with regard to standard prognostic factors in other studies. It may also be useful to compare outcome for the selected patients (M+) to the outcome for the patients of the same series who did not have such predicted good prognosis (M-). The latter patients would have received chemotherapy, but their outcome even with chemotherapy may not be as good as the M+ patients without chemotherapy. If that is the case, then the value of the marker for withholding chemotherapy will have been demonstrated.

An alternative approach would be to randomize patients selected as low risk based on marker status (M+) to either receive or not receive chemotherapy. The marker would be validated if the randomized trial demonstrated that there was no clinically significant benefit of chemotherapy in the selected subset of patients. This would have to be a very large clinical trial, however. The benefit of chemotherapy would only be expected to reduce the hazard or recurrence by approximately 25% and with a very low event rate this is equivalent to a very small difference in absolute disease-free survival. The randomized trial asks a different question than the strategy described in the previous paragraph. The randomized trial asks whether there is a benefit of treatment. For very good prognosis patients, however, a statistically significant treatment effect may be of questionable clinical significance. Consequently, the randomized trial may not answer the most relevant question.

Table 2 shows criteria of Gasparini et al. [11] for the adoption into clinical practice of new prognostic markers for use in treatment selection for patients with node negative breast cancer.

4.2 Identifying Node Positive Patients Who Do Not Benefit From A Chemotherapy Regimen

Consider now a putative marker that permits the identification of patients who do not benefit from a chemotherapy regimen which has been standard treatment. Let T denote the chemotherapy regimen, and let S denote local therapy or local therapy plus

Tamoxifen. The retrospective study used to develop the marker may have been based on tissue from a randomized trial of T vs S. In some cases the marker may be based on finding a signature of patients who do not respond to T in metastatic disease trials.

The ideal validation trial would probably be a randomized trial of S versus T for patients with node positive breast cancer. One could analyze such a trial by seeing whether the benefit of T versus S depended on the marker level. Such a trial would generally be impractical, however, because T or some other kind of chemotherapy is standard treatment for node positive patients. It might be possible, however, to randomize patients to receive or not receive one or more courses of T pre-operatively, and to correlate marker result with biological response to T as assessed from the surgical specimen.

A second strategy would be to use chemotherapy T on all patients after measuring the marker. One could then determine prospectively whether the marker level correlates with outcome. This is a strategy analogous to that recommended in section 4.1. Here, however, one is trying to determine whether the marker identifies a group of such poor disease-free-survival on standard treatment T that the chemotherapy is judged non-worthwhile even in the absence of a control group not receiving chemotherapy. This strategy may be less satisfactory for judging poor prognosis in absolute terms than it was in 4.1 for judging good prognosis.

A third strategy would be to randomize the patients to marker based versus non-marker based therapeutic management. The non-marker based management would assign T to all patients. The marker based management would assign T to all except those predicted based on the marker to be non-responsive (M-). One way of conducting such a trial is to measure the marker only for those patients assigned to marker based management. The value of the marker is determined by measuring the outcome for the marker based management arm to the outcome for the non-marker based management arm. This is, however, a very inefficient trial design. Because most patients in both arms of the trial will be receiving the same treatment, the average treatment difference will be very small between the arms and a huge sample size will be required. The situation is even more problematic because it is a therapeutic equivalence trial in the sense that failure to find a statistically significant difference leads to the adoption of the new treatment approach, in this case marker based treatment assignment.

A better design is to measure the marker on all patients, and then randomize them to marker based treatment versus non-marker based treatment. The evaluation of the marker can be performed by comparing outcomes for the M- patients who received chemotherapy T on the non-marker based arm but treatment S on the marker based arm. This will require a much smaller sample size than the design of the previous paragraph. This design is essentially equivalent to randomizing the M- patients to T or S.

4.3 Identifying Node Positive Patients Who Benefit From a Specific Regimen

Our third scenario is that we have a putative marker that identifies patients whose tumors are responsive to a new regimen E when the standard chemotherapeutic regimen is T. Many new therapeutics have defined molecular targets and are developed in conjunction with an assay that measures the expression of the target. The most adequate validation study is often a randomized clinical trial in which both marker positive and marker negative patients are randomized to either standard treatment T or T plus the new regimen E. The trial should be large enough so that the new regimen can be evaluated separately in the M+ and M- subsets. This requires about twice as many patients as if the regimen T+E were to be evaluated overall, without reference to the marker.

If the biological relationship between the marker and the therapeutic is sufficiently strong, it may be difficult to justify including marker negative patients in the study. A randomized study comparing T to T+E for M+ patients may be very efficient for demonstrating the effectiveness of the new treatment E, but it will not really constitute a validation of the marker. The development of the therapeutic, supported by the marker assay, may, however, be more important than validation of the essentiality of the marker for selecting patients.

The least desirable alternative would be to randomize patients between T and T+E without measuring the marker. If the marker is important, then such a trial design may be very inefficient for evaluating the therapeutic E, and of course, it provides no information for validating the marker.

The scenario described here is also applicable to the development of treatment regimens in which the molecular target is not known or not known with certainty. Instead of using an assay based on the expression of the putative target, one may use a DNA microarray based classifier developed in phase II trials of metastatic disease patients for distinguishing responders from non-responders to the new regimen E. If tumor specimens are available from patients treated with the standard treatment T as well as those treated with the new treatment E, the classifier can be developed to identify patients who are predicted to be more responsive to the new treatment E but not to standard treatment T.

Table 1: Guidelines for Validation Studies [10]

1. Intra- and inter-laboratory reproducibility of assays should be documented
2. Laboratory assays should be performed blinded to clinical data and outcome
3. An inception cohort of patients should be assembled with <15% of patients non-evaluable due to missing tissue or data. The referral pattern and eligibility criteria should be described
4. Treatment should be standardized or randomized and accounted for in the analysis
5. Hypotheses should be stated in advance, including specification of prognostic factors, coding of prognostic factors, endpoints, and subsets of patients and treatments
6. The sample size and number of events should be sufficiently large that statistically reliable results are obtained. Statistical power calculations that incorporate the number of hypotheses to be tested and appropriate subsets for each hypothesis should be described. There should be at least 10 events per prognostic factor examined per subset analyzed.
7. Analyses should test whether new factors add predictiveness after adjustment for or within subsets determined by standard prognostic factors
8. Analyses should be adjusted for the number of hypotheses to be tested
9. Analyses should be based on pre-specified cutoff values for prognostic factors or cutoffs should be avoided

Table 2 : Guidelines for Introduction of New Biomarker for Node Negative Breast Cancer into Clinical Practice [11].

1. Favorable previous steps in at least two independent studies
2. Favorable cost/benefit ratio
3. Availability of a feasible, reproducible, and sensitive method to detect the indicator
4. Identification of the most appropriate adjuvant therapy for the high-risk subgroup of node-negative patients in relation to the indicator used
5. A proven advantage for the treated node-negative high-risk patients of at least 20%-25% in relapse-free survival or overall survival versus untreated node-negative control subjects by at least two independent comparative trials

1. Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? Breast Cancer Research and Treatment 1992; 22:197-206.
2. Simon R. Using DNA microarrays for diagnostic and prognostic prediction. Expert Review of Molecular Diagnostics 2003; (In Press).
3. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19:368-375.
4. Kattan MW. Judging new markers by their ability to improve predictive accuracy. Journal of the National Cancer Institute 2003; 95:634-635.
5. Korn EL, Simon R. Measures of explained variation for survival data. Statistics in Medicine 1990; 9:487-504.
6. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: Class prediction methods. Journal of the National Cancer Institute 2003; 95:14-18.
7. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 2002; 346:1937-1947.
8. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 2002; 9:505-511.
9. Vasselli J, Shih JH, Iyengar SR, et al. Predicting survival in patients with metastatic kidney cancer by gene expression profiling in the primary tumor. Proceedings of the National Academy of Science U.S.A. 2003; 100:6958-6963.
10. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. British Journal of Cancer 1994; 69:979-985.
11. Gasparini G, Pozza F, Harris AL. Evaluating the potential usefulness of new prognostic and predictive indicators in node-negative breast cancer patients. Journal of the National Cancer Institute 1993; 85:1206-1219.