Use of Predictive Biomarker Classifiers in the Design of Pivotal Clinical Trials

Richard Simon

Key words: Predictive biomarker, clinical trial design, gene expression, supervised classification, targeted clinical trial

Richard Simon, D.Sc.
Biometric Research Branch
National Cancer Institute
9000 Rockville Pike
MSC 7434
Bethesda MD 20892-7434
Tel 301.496-0975
Fax 301.402-0560
rsimon@nih.gov

Summary

In this chapter we distinguish the use of predictive biomarkers from surrogate endpoint biomarkers. We also distinguish the use of predictive biomarkers for selecting patients for pivotal clinical trials of a new drug from the use of predictive biomarkers for optimizing the utilization of an existing drug. We summarize the key steps in the development of predictive biomarker classifiers for use in new drug development. We discuss the design of targeted clinical trials in which a predictive biomarker classifier is used to restrict entry and present results comparing the efficiency of targeted trials relative to standard randomized pivotal trials. We also discuss alternative designs in which the predictive biomarker classifier is not used to restrict entry of patients but is used to prospectively define an analysis plan for evaluated the new drug in classifier negative and positive patients. The development of predictive biomarker classifiers can be subjective, but pivotal trials should test hypotheses about the effectiveness of a new drug in subsets defined in a completely pre-specified manner by a predictive classifier and should not contain any subjective components. The data used to develop the predictive classifier should be distinct from the data used to evaluate a new drug in subsets determined by the classifier. The purpose of the pivotal trial is to evaluate the new drug in patient groups defined prospectively by the predictive classifier, not to refine or re-evaluate the classifier or its components. New drug development should move from a correlative science mode to a predictive medicine mode.

1. What is a Predictive Biomarker?

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 there is also interest in using and endpoint biomarker in pivotal 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 [1-3]. These standards are stringent because of the key role of the pivotal trial endpoint in claims. 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 pivotal totals using standard measures of clinical outcomes as and endpoint.

Biomarkers can be pre-treatment measurements used to characterize the patient's disease in order to determine whether the patient is a good candidate for a treatment. These are called *predictive biomarkers*. 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 the heterogeneously treated patients. Most prognostic factor studies are based on convenience samples of patients for whom tissue is available. The studies are often not focused on a particular medical decision facing physicians and hence the resulting prognostic actors identified have no therapeutic relevance and are not widely used. 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

2. Development 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, for example, presence or absence of a point mutation in the EGFR gene, or amplification of the HER2 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. The individual gene expression levels must be combined in some mathematically determined manner.

There are two kinds of gene expression based classifiers that are frequently used. The first is based on a weighted average of expression with tumor response or patient outcome. A training set of data is used 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. A weighted average of the expression levels for the signature genes is adopted as a predictive index. Many of the commonly used classifier types are based on such weighted averages. These include Golub's weighted voting classifier [4], the combined covariate predictor [5], Fishers linear discrimant and diagonal linear discrimant analysis [6], support vector markers with inner product kernel [7], naive Bayes classifier [8], and perceptrons [9]. 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 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 [10]. Unfortunately, cross validation methods are used improperly in many cases [11].

The second kind of binary classifier widely used for gene expression data are the non-parametric distance based methods including nearest neighbor, k-nearest neighbor, nearest centroid and shrunken centroid classifiers [6, 12]. These methods also use signature genes selected based on the correlation of their expression levels with response or outcome. 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 distance 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 (smallest distance). 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 are similar. The comments made in the previous paragraph about use of independent data to evaluate prediction accuracy apply equally to these non-parametric distance based classifiers.

Although many other types of binary classifiers have been developed and strong claims for them are often made by their developers, independent evaluation have generally concluded that other more complex methods rarely outperform weighted average based methods or non-parametric distance based methods. For any training set of data it is recommended here to develop weighted average based classifiers and use either complete cross-validation or a separate test set of data to evaluate prediction accuracy of these methods and select one for use n pivoted trials. The BRB-Array Tools Software [13] provides a convenient integrated environment for identifying signature genes, developing weighted average and non-parametric distance based classifiers, and validly evaluating prediction accuracy. 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 [14-16].

3. Use of Predictive Biomarkers in the Design of Pivotal Trials

The objective of a pivotal clinical trial is to evaluate a new drug, given in a defined manner, has a medical utility for a defined hypothesis about treatment effectiveness in specified patient population groups, they are not exploratory laboratories. The role of a predictive biomarker classifier is to specify 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 pivotal trial is to be used to

develop or refine a biomarker classifier, then treatment hypotheses involving that classifier should be tested in a separate pivotal trial. One exception is the adaptive trial o the Friedlin and Simon [17] where some data from a pivotal trial is used to develop a classifier and that data is excluded from the data for that same pivotal trial that is used to test a treatment hypothesis in the subset of patients defined as positive by that classifier.

Figure 1 depicts the process of developing a predictive biomarker classifier and using it to restrict eligibility to a pivotal trail. The purpose of the trial is to evaluate the new treatment regiment in the patients defined as classifier positive by the classifier. The purpose is not to re-evaluate, redevelop or refine the classifier. The purpose of the study is to evaluate the new treatment regimen in classifier positive patients, not to validate the predictiveness of the classifier. If the treatment is shown to be effective there is a reproducible assay for classifier positively, then there is a medical utility for approval of the treatment in classifier positive patients even if the treatment hasn't been tested in classifier negative patients. In cases where the classifier is biologically based on the target of the drug, it may not be ethically appropriate to treat classifier negative patients. Even where the classifier is empirically based it is questionable whether registration of a treatment for a set of patients in whom medical utility has been established should be contingent on evaluation on evaluation of the treatment in a different set of patients. Whether the trial design shown in Figure 1 is sufficient for licensing the classifier itself is somewhat more complex and depends on the regulatory agency, specific regulatory language and agency interpretations. A classifier for which a reproducible assay exists itself would seem to have medical utility if it identifies a set of patients.

4. Efficiency of Targeted Designs

Simon and Maitournam [18-20] evaluated the efficiency of the targeted design shown in Figure 1 relative to the conventional broad randomization design in which the classifier is not used to restrict entry. One measure of relative efficiency is the number of randomized patients required for the targeted design relative to the conventional design. A second measure of efficiency is the number of patients required for screening in the targeted design relative to the number required to randomize for the standard design. If n_T randomized patients are required for the targeted design and γ_+ is the fraction of patients who are classifier positive, then approximately n_T / γ_+ patients are required to be screened for the targeted design. For example, suppose half as many patients are required for randomization with the targeted design and compared to the standard design but only 25% of the patients are classifier positive. The twice as many patients will be required for screening for the targeted design that for randomization with the standard design.

Relative efficiency of the targeted and standard designs depends on the specificity of benefit of the new treatment for classifier positive patients and the prevalence of the classifier positivity. The specificity of the treatment benefit can itself be decomposed into specificity of treatment benefit for the biological state measured by the assay and measurement accuracy of the assay. For example, the biological state may be an amplification of a gene and it is possible that treatment benefit for classifier negative patient results both from some treatment benefit in patients without the amplification and

from the false negative assays for gene amplification. Usually there will not be separate estimates for these components of treatment specificity and hence no real value for considering them separately in planning a pivotal trial. Here, we will use the composite effect.

Simon and Maitournam [18, 21] showed that for binary endpoint trials that the ratio of number of patients required for the randomization in the standardized trial compared to the targeted trial is approximately

$$n_S / n_T \approx \left(\frac{1}{\gamma_+ + (1-\gamma_+)\delta_- / \delta_+} \right)^2 f \qquad (1)$$

where γ_+ denotes the proportion classifier positive and δ_- / δ_+ is the ratio of the treatment effect for classifier negative patients to the treatment effect for classifier positive patients. The parameter f is generally close to 1 unless the control response ratio is very low. In cases where benefit of the new treatment is limited to classifier positive patients, $\delta_- = 0$ and the formula simplifies to f / γ_+^2. If the treatment is half as effective in classifier negative patients as classifier positive patients then the formula simplifies to $4f / (\gamma_+ + 1)^2$. Table 1 shows the ratio of the number of randomized patients using the formula with f=1.

Since the number of patients required to screen for the targeted trial is n_T / γ_+, the screened ratio of efficiency is

$$n_S / screened_T \approx \frac{\gamma_+ f}{\left(\gamma_+ + (1-\gamma_+)\delta_- / \delta_+\right)^2} \qquad (2)$$

If $\delta_- = 0$ this equals f / γ_+. If the treatment is half as effective for classifier negative patients as for classifier positive patients then (2) equals $4\gamma_+ f / \left(\gamma_+ + 1\right)^2$. The screened ratio approximate efficiency for these two cases is also illustrated in Table 1.

When the proportion of classifier positive patients is less than one half, the number of patients required for randomization in the targeted design is much smaller than for the standard design, at least by a factor of two, regardless of whether the treatment effect is completely specific for classifier positive patients or whether the classification negative patients benefit half as much as the positive patients. In the former case, however, the targeted design also requires many fewer patients to screen than required for randomization with the standard design. If, however, the treatment effect for classifier negative patients is half that for the classifier positive patients, then the targeted design may require more patients to screen that are required for randomization with the standard design. Hence, this targeted design is most appropriate when the treatment benefit is expected to be quite specific for classifier positive patients. When the proportion of patients who are classifier positive exceeds 50%, the efficiency advantages of the targeted trial are reduced.

A web based interactive program for planning targeted clinical trials is available at http://linus.nci.nih.gov/brb It provides a comparison of the targeted design to standard design with regard to number of randomized and screened patients. It uses more accurate formulas than the approximations utilized above and also provides the comparison for studies in which there is a time-to-event endpoint such as survival or progression-free survival. Figure 2 shows a screen shot of the web page of input dialog for the time-to-event calculation. For the example shown, the median survival for the control group is 1 year and 25% of the patients are classifier positive. For power calculations it is postulated that the new treatment reduces the hazard of death by 50% for the classifier positive patients and is ineffective for the classifier negative patients. A screen shot of the output of the program is shown in Figure 3. With accrual of 100 patients per year and a follow-up period of two years after end of accrual, a targeted trial of 4.27 years would randomize 107 target positive patients and achieve power 0.90. In contrast, an untargeted trial of 4.27 years of accrual would randomize 427 patients but have statistical power only 0.45 because the overall treatment effect is so diluted by the lack of treatment effect in the 75% of patients who are classifier negative. In this setting the targeted trial is very advantageous. If however, the treatment reduced the hazard of death by 20% for classifier negative patients, then the statistical power of the untargeted design after 4.27 years of accrual is 0.925, very similar to that of the targeted design (results not shown). In that circumstance, the targeted design is not advantageous. The targeted design is most valuable when treatment benefit is limited to target positive patients and the assay for measuring the classifier is quite accurate.

5. Stratified designs

Simon and Wang [22]have described clinical trial designs in which both classifier negative and classifier positive patients are randomized and the classifier is measured. In this case, it is important to have a pre-defined analysis plan for using the classifier information. It is not sufficient to merely "stratify" the randomization process by the classifier. Simon and Wang propose dividing the usual 5% type I error into a portion $\alpha_{overall}$ for comparing the treatment groups overall for all randomized patients and a portion $.05-\alpha_{overall}$ for comparing the treatment groups in the pre-defined subset of patients who are classifier positive. The subset test would only be performed if the overall test is not significant at the reduced threshold $\alpha_{overall}$. A web based interactive program for planning stratified clinical trials of this type is also available at http://linus.nci.nih.gov/brb .

Figure 1

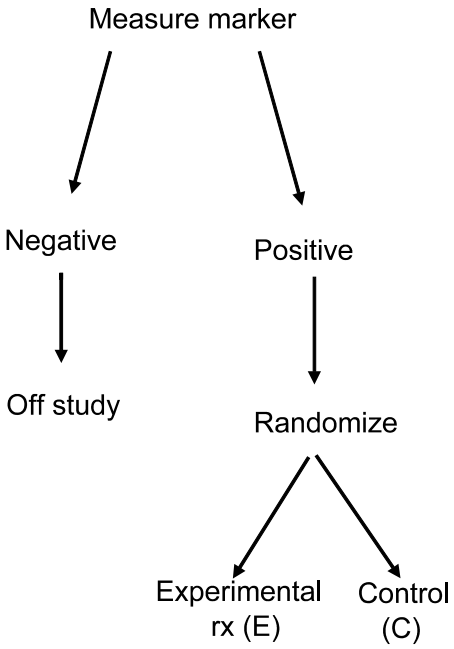
Measure marker
Negative
Positive
Off study
Randomize
Experimental rx (E)
Control (C)

Figure 2

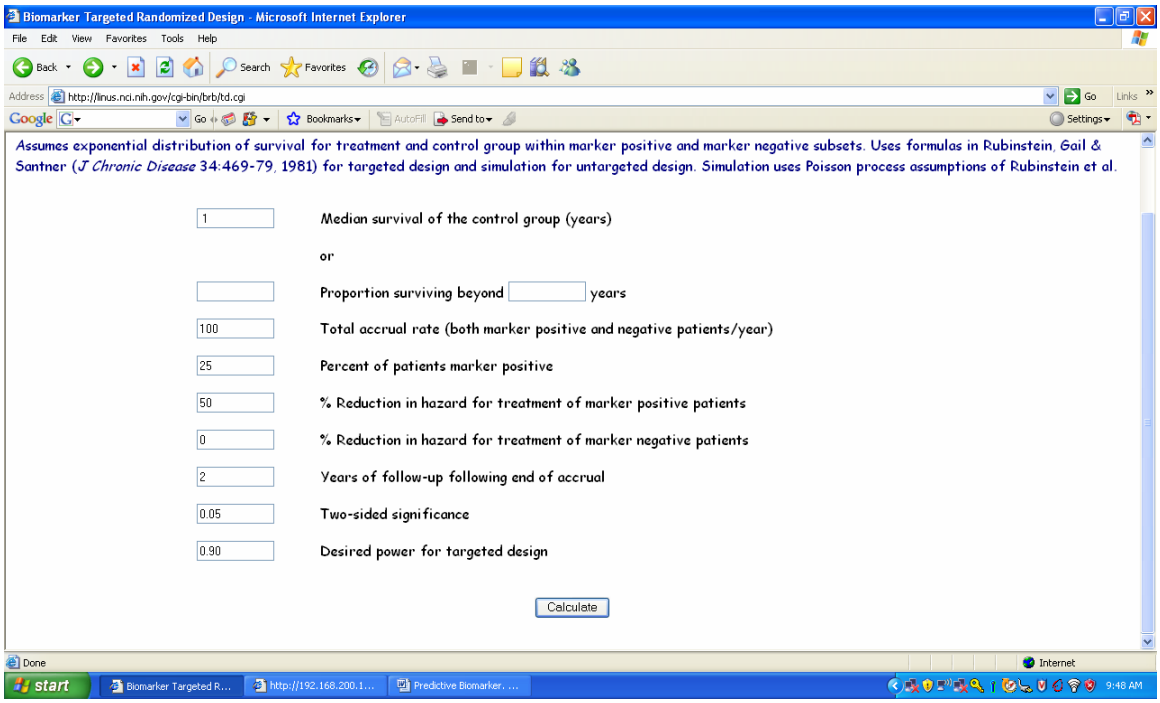
Biomarker Targeted Randomized Design - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://linus.nci.nih.gov/cgi-bin/brb/td.cgi
Go Links
Google Go Bookmarks AutoFill Send to Settings
Assumes exponential distribution of survival for treatment and control group within marker positive and marker negative subsets. Uses formulas in Rubinstein, Gail & Santner (J Chronic Disease 34:469-79, 1981) for targeted design and simulation for untargeted design. Simulation uses Poisson process assumptions of Rubinstein et al.
1
Median survival of the control group (years)
or
Proportion surviving beyond years
100
Total accrual rate (both marker positive and negative patients/year)
25
Percent of patients marker positive
50
% Reduction in hazard for treatment of marker positive patients
0
% Reduction in hazard for treatment of marker negative patients
2
Years of follow-up following end of accrual
0.05
Two-sided significance
0.90
Desired power for targeted design
Calculate
Done
Internet
start
Biomarker Targeted R...
http://192.168.200.1...
Predictive Biomarker. ...
9:48 AM

Figure 3

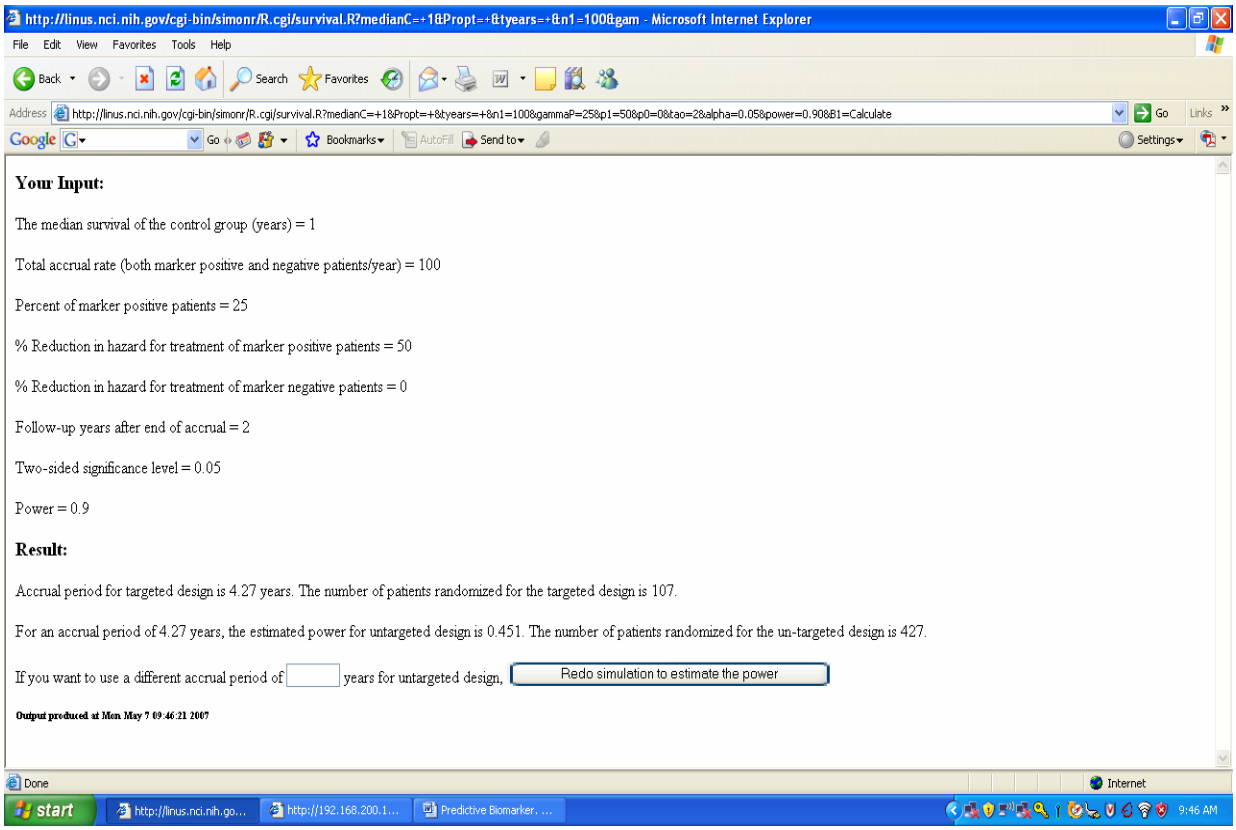
http://linus.nci.nih.gov/cgi-bin/simonr/R.cgi/survival.R?medianC=+1&Propt=+&tyears=+&n1=100&gam - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://linus.nci.nih.gov/cgi-bin/simonr/R.cgi/survival.R?medianC=+1&Propt=+&tyears=+&n1=100&gammaP=25&p1=50&p0=0&tao=2&alpha=0.05&power=0.90&B1=Calculate
Go Links
Google Go Bookmarks AutoFill Send to Settings
Your Input:
The median survival of the control group (years) = 1
Total accrual rate (both marker positive and negative patients/year) = 100
Percent of marker positive patients = 25
% Reduction in hazard for treatment of marker positive patients = 50
% Reduction in hazard for treatment of marker negative patients = 0
Follow-up years after end of accrual = 2
Two-sided significance level = 0.05
Power = 0.9
Result:
Accrual period for targeted design is 4.27 years. The number of patients randomized for the targeted design is 107.
For an accrual period of 4.27 years, the estimated power for untargeted design is 0.451. The number of patients randomized for the un-targeted design is 427.
If you want to use a different accrual period of years for untargeted design, Redo simulation to estimate the power
Output produced at Mon May 7 09:46:21 2007
Done
Internet
start
http://linus.nci.nih.go...
http://192.168.200.1...
Predictive Biomarker, ...
9:46 AM

Table 1

Proportion Classifier Positive	$\delta_- / \delta_+ = 0$		$\delta_- / \delta_+ = .5$	
	Randomized for Standard Design / Randomized for Targeted Design	Randomized for Standard Design / Screened for Targeted Design	Randomized for Standard Design / Randomized for Targeted Design	Randomized for Standard Design / Screened for Targeted Design
.5	4	2	1.8	0.89
.4	6.25	2.5	2.0	0.82
.3	11.1	3	2.4	0.71
.2	25	5	2.8	0.56
.1	100	10	3.3	0.33

REFERENCES

1. 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.
2. 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.
3. Korn EL, Albert PS, McShane LM. Assessing surrogates as trial endpoints using mixed models. Statistics in Medicine 2004;24:163-82.
4. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression modeling. Science 1999;286:531-7.
5. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 2002;9:505-11.
6. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York: Springer; 2003.
7. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Science USA 2001;98(26):15149-54.
8. Hand DJ, Yu K. Idiot's Bayes-Not so stupid after all? International Statistical Review 2001;69(3):385-98.
9. Khan J, Wei JS, ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 2001;7(6):673-9.
10. 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;95:14-8.
11. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. (Submitted for publication) 2007;99:147-57.
12. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Science USA 2002;99(10):6567-72.
13. 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.
14. Simon R. Using DNA microarrays for diagnostic and prognostic prediction. Expert Review of Molecular Diagnostics 2003;3(5):587-95.
15. Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. British Journal of Cancer 2003;89:1599-604.
16. Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology 2005;23:7332-41.
17. 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.

18. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 2004;10:6759-63.
19. Simon R, Maitournam A. Correction & Supplement: Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 2006;12:3229.
20. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine 2005;24:329-39.
21. Simon R, Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials: Supplement and Correction. Clinical Cancer Research 2006;12:3229.
22. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. The Pharmacogenomcs Journal 2006;6:1667-173.