Cancer Biomarkers 2 (2006) 89–96
IOS Press

Validation of pharmacogenomic biomarker classifiers for treatment selection

Richard Simon*
Biometric Research Branch, NIH, Bethesda, MD, USA

Abstract. Physicians need improved tools for selecting treatments for individual patients. Many syndromes traditionally viewed as individual diseases are heterogeneous in molecular pathogenesis and treatment responsiveness. This results in treatment of many patients with ineffective drugs and leads to the conduct of large clinical trials to identify small average treatment benefits for heterogeneous groups of patients. New genomic and proteomic technologies provide powerful tools for the selection of patients likely to benefit from a therapeutic without unacceptable adverse events. In spite of the large literature on developing predictive biomarkers and on statistical methodology for analysis of high dimensional data, there is considerable uncertainty about the validation of biomarker based diagnostic classifiers for treatment selection. In this paper we attempt to clarify these issues and to provide guidance on the design of clinical trials for evaluating the clinical utility and robustness of pharmacogenomic classifiers.

Keywords: Pharmacogenomics, biomarker, genomics, DNA microarray, clinical trial design, validation

1. Introduction

Physicians need improved tools for selecting treatments for individual patients. For example, many cancer treatments benefit only a minority of the patients to whom they are administered. Being able to predict which patients are most likely to benefit would not only save patients from unnecessary toxicity and inconvenience, but might facilitate their receiving drugs that are more likely to help them. In addition, the current over-treatment of patients results in major expense for individuals and society, an expense which may not be indefinitely sustainable. In this paper we will address some key issues in the validation of pharmacogenomic classifiers.

2. Pharmacogenomic classifiers

Much of the discussion about disease biomarkers is in the context of markers which measure some aspect of disease status, extent, or activity. Such biomarkers are often proposed for use in early detection of disease or as a surrogate endpoint for evaluating prevention or therapeutic interventions. The validation of such biomarkers is difficult for a variety of reasons, but particularly because the molecular pathogenesis of many diseases is incompletely understood and hence it is not possible to establish the biological relevance of a measure of disease status.

A pharmacogenomic biomarker is any measurable quantity that can be used to select treatment; for example, the result of an immunohistochemical assay for a single protein, the abundance of a protein in serum, the abundance of mRNA transcripts for a gene in a sample of disease tissue or the presence/absence status of a specified germline polymorphism or tumor mutation. A pharmacogenomic classifier is a mathematical function that translates the biomarker values to a set of prognostic categories. These categories generally correspond to levels of predicted clinical outcome. With the advent of gene expression profiling, it is increasingly common to define composite pharmacogenomic classifiers based on the levels of expression of dozens of genes. For a fully specified classifier, however, all of the parameters and cut-points are specified for de-

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

ISSN 1574-0153/06/$17.00

termining how to weight the different components and how to map the multivariate data into a defined set of categories. A completely defined classifier can be used to select patients and stratify patients for therapy in clinical trials that enable the clinical value of the classifier to be evaluated. Specifying only the genes involved does not enable one to structure prospective clinical validation experiments in which patients are assigned or stratified in prospectively well defined ways. Repeatedly correlating expression of individual genes against outcomes does not constitute adequate evaluation of the medical value of a diagnostic technology for treatment selection.

3. Validation

Biomarkers are used for very different purposes and validation should relate to fitness for a defined purpose. It is not likely to be productive to require validation in some more absolute biological sense for diseases whose molecular pathogenesis is not fully understood. For pharmacogenomic biomarkers, we will focus on the design of validation studies to establish clinical benefit in assisting with treatment selection. For example, the Oncotype-Dx risk score was developed to measure prognosis for node negative, estrogen receptor positive patients with primary breast cancer receiving Tamoxifen therapy after surgical resection of the primary lesion [6]. The validation issue is whether use of this risk score results in clinical benefit. The components of expression signature classifiers need not be valid biomarkers in the sense of the Food and Drug Administration [1]. Those criteria require that the role of the biomarker be mechanistically understood and accepted as markers of disease activity. Such criteria are relevant for biomarkers used as surrogate endpoints but not for the components of expression signatures used for tailoring treatments. It is, of course, desirable to understand the mechanistic relationship of the components of an expression signature, but the classifier can be validated without such understanding.

4. Developmental and validation studies

It is important to distinguish the studies which develop parmacogenomic classifiers from those which evaluate the clinical utility of such classifiers. The vast majority of published prognostic marker studies are developmental and are not adequate for establishing the clinical utility and robustness of a classifier [11]. Developmental studies are often based on a convenience sample of patients for whom tissue is available but who are heterogeneous with regard to treatment and stage. The studies are generally performed in an exploratory manner with no written protocol, no eligibility criteria, no primary endpoint or hypotheses and no defined analysis plan. The analysis often includes numerous analyses of different endpoints and patient subsets. Often there are multiple candidate biomarkers to evaluate and multiple ways of measuring and combining the candidate biomarkers. Such an informal approach is appropriate in a developmental study so long as one recognizes that the same study cannot be used to evaluate the clinical value of the resulting biomarkers or classifiers. The developmental study is exploratory and directed to hypothesis formation. The special statistical issues involved in development of genomic classifiers based on high dimensional data are a topic in themselves. Some of these issues are reviewed elsewhere (e.g. [15]).

Developmental studies should develop completely specified classifiers and completely specified hypotheses that can be tested in subsequent validation studies. Although there is a large literature on prognostic markers, few such factors are used in clinical practice. To a large extent this is due to a lack of adequate validation studies which demonstrate the therapeutic relevance and robustness of pre-specified biomarker classifiers. Prognostic markers are unlikely to be used unless they are therapeutically relevant and most developmental studies, unless they are based on patients treated in a single clinical trial, are not based on a cohort medically coherent enough to establish therapeutic relevance. Developmental studies rarely establish the robustness of the classifier and of the underlying assays under conditions that simulate those likely to be found in real world patient management.

4.1. Estimates of predictive accuracy based on developmental studies

Developmental studies are analogous to phase 2 clinical trials. They should include an indication of whether the genomic classifier is promising and worthy of phase 3 evaluation. There are special problems in evaluating whether classifiers based on high dimensional genomic or proteomic assays are promising however. The difficulty derives from the fact that the number of candidate features available for use in the classifier is much larger than the number of cases available for analysis. In such situations, it is always possible to find classifiers that

accurately classify the data on which they were developed even if there is no relationship between expression of any of the genes and outcome [8]. Consequently, even in developmental studies, some kind of validation on data not used for developing the model is necessary. This "internal validation" is usually accomplished either by splitting the data into two portions, one used for training the model and the other for testing the model, or some form of cross-validation based on repeated model development and testing on random data partitions. This internal validation should not, however, be confused with external validation of the classifier utility in a setting simulating broad clinical application.

The most straightforward method of estimating the prediction accuracy is the *split-sample* method of partitioning the set of samples into a training set and a test set. Rosenwald et al. [9] used this approach successfully in their international study of prognostic prediction for large B 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 single fully specified prediction model, they accessed the test set for the first time. On the test set there was no adjustment of the model 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. In addition to estimating the overall error rate on the test set, one can also estimate other important operating characteristics of the test such as sensitivity, specificity, positive and negative predictive values.

The split-sample method is often used with so few samples in the test set, however, that the validation is almost meaningless. One can evaluate the adequacy of the size of the test set by computing the statistical significance of the classification error rate on the test set or by computing a confidence interval for the test set error rate. Since the test set is separate from the training set, the number of errors on the test set has a binomial distribution.

Michiels et al. [4] suggested that multiple training-test partitions be used, rather than just one. The split sample approach is mostly useful, however, when one does not have a well defined algorithm for developing the classifier. When there is a single training set-test set partition, one can perform numerous unplanned analyses on the training set to develop a classifier and then test that classifier on the test set. With multiple training-test partitions however, that type of flexible approach to model development cannot be used. If one has an algorithm for classifier development, it is generally better to use one of the cross-validation or bootstrap resampling approaches to estimating error rate (see below) because the split sample approach does not provide as efficient a use of the available data [5].

Cross-validation is an alternative to the split sample method of estimating prediction accuracy [8]. Molinaro et al. describe and evaluate many variants of cross-validation and bootstrap re-sampling for classification problems where the number of candidate predictors vastly exceeds the number of cases [5]. The cross-validated prediction error is an estimate of the prediction error associated with application of the algorithm for model building to the entire dataset.

A commonly used invalid estimate is called the *re-substitution estimate*. You use all the samples to develop a model. Then you predict the class of each sample using that model. The predicted class labels are compared to the true class labels and the errors are totaled. It is well-known that the re-substitution estimate of error is biased for small data sets and the simulation of Simon et al. [14] confirmed that, with an astounding 98.2% of the simulated data sets resulting in zero misclassifications even when no true underlying difference existed between the two groups.

Simon et al. [14] also showed that cross-validating the prediction rule after selection of differentially expressed genes from the full data set does little to correct the bias of the re-substitution estimator: 90.2% of simulated data sets with no true relationship between expression data and class still result in zero misclassifications. When feature selection was also re-done in each cross-validated training set, however, appropriate estimates of mis-classification error were obtained; the median estimated misclassion rate was approximately 50%.

The simulation results underscore the importance of cross-validating all steps of predictor construction in estimating the error rate. It can also be useful to compute the statistical significance of the cross-validated estimate of classification error. This determines the probability of obtaining a cross-validated classification error as small as actually achieved if there were no relationship between the expression data and class identifiers. A flexible method for computing this statistical significance was described by Radmacher et al. [8]. It involves randomly permuting the class identifiers among the patients and then re-calculating the cross-validated classification error for the permuted data. This is done a large number of times to generate the null distribution of the cross-validated prediction error. If the value

of the cross-validated error obtained for the real data lies far enough in the tail of this null distribution, then the results are statistically significant. This method of computing statistical significance of cross-validated error rate for a wide variety of classifier functions is implemented in the BRB-ArrayTools software [12]. Statistical significance, however, does not imply that the prediction accuracy is sufficient for the test to have clinical utility.

Even if a classifier is developed for a set of patients sufficiently homogeneous and uniformly treated to be therapeutically relevant, it may be important to evaluate whether the classifier predicts more accurately than do standard prognostic factors or adds predictive accuracy to that provided by standard prognostic factors. For example, Rosenwald et al. [9] developed a classifier of outcome for patients with advanced diffuse large B cell lymphoma receiving CHOP chemotherapy. The International Prognostic Index (IPI) is easily measured and prognostically important for such patients, however, and so it was important for Rosenwald et al. to address whether their classifier provided added value. The most effective way of addressing whether a classifier adds predictive accuracy to a standard classification system is to examine outcome for the new system within the levels of the standard system.

5. Design of validation studies

The objective of external validation is to determine whether use of a completely specified diagnostic classifier for therapeutic decision making in a defined clinical context results in patient benefit. Patient benefit may represent better efficacy, reduced incidence of adverse events, better convenience or lower costs. The objective is not to repeat the developmental study and see if the same genes are prognostic or if the same classifier is obtained.

An independent validation study could be a prospective clinical trial in which patients are randomized to treatment assignment without use of the classifier versus treatment assignment with the aid of the classifier. This design requires that the classifier be determined only in half of the patients. Often, however, this design will be inefficient and require a huge sample size because many or most of the patients will receive the same treatment either way they are randomized. For example, consider women with lymph node negative, ER positive breast cancers. About one third of such patients might be expected to be classified as low risk for recurrence based on the Oncotype-DX expression signature based risk score [6]. If one wants to test the strategy of withholding cytotoxic chemotherapy (systemic treatment with Tamoxifen alone) from the subset of patients classified as low risk, it would be inefficient to randomize all of the node negative ER positive patients. If one randomizes all the patients and only performs the assay on the half randomized to have classifier based therapy, then the two randomization groups must be compared overall, although two thirds of the patients receive the same treatment in both arms. The classifier is considered clinically useful if the outcome for the two randomized groups are equal because some of the patients in the group with marker determined therapy were spared the side effects of chemotherapy. This is a therapeutic equivalence design, but a very problematic one since most patients in both randomization arms receive the same therapy.

Table 1
Approximate number of events required for 80% power with 5% two-sided log-rank test for comparing arms of design shown in Fig. 1. Randomized arms are mixtures of marker – and marker + patients. Hazard ratio for marker – patients is 1 for the two treatment groups and 0.67 for marker + patients. All patients are followed to failure

Proportion of patients marker +	Approximate number of events required
20%	5200
33%	1878
50%	820

Table 1 indicates the approximate number of events required for the "Assay after randomization" design of Fig. 1. Table 1 is based on the assumption that time-to-event distributions are exponential, and that the hazard ratio between treatment groups is 0.67 in marker positive patients and 1.0 in marker negative patients. For application to the breast cancer example considered above, recall that the marker negative patients receive the same treatment with marker determined treatment or standard of care treatment. The number events required to obtain 80% statistical power for detecting statistical significance using a 5% two-sided log-rank test are shown in Table 1 in terms of the proportion of patients who are marker positive. If only one third of the patients are marker positive, then approximately 1878 events are required. If the average event rate over the follow-up period is 10%, then observing 1878 events requires the accrual of 18,780 patients. Designs related to that shown in Fig. 1 have been discussed by Sargent et al. [10].

A more efficient alternative is to perform the assay up front for all patients, and then randomize only those classified as low risk. Those patients would be

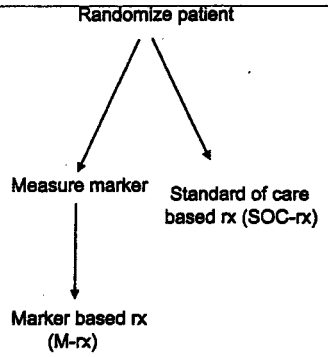

Fig. 1. Randomized clinical trial for evaluating whether use of a biomarker based classifier for treatment selection results in improved clinical outcome. All patients with conventional diagnosis are randomized between biomarker based treatment (M-rx) or standard of care based treatment (SOC-rx). This design is often very inefficient.

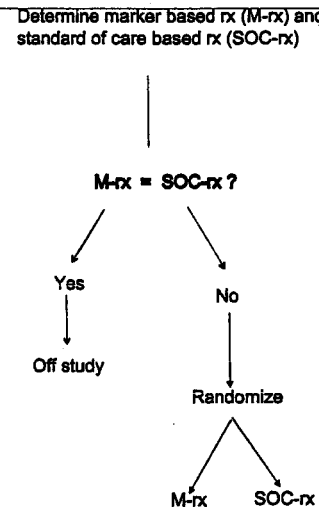

Fig. 2. Improved clinical trial design for evaluating whether use of a biomarker based classifier for treatment selection results in improved clinical outcome. The biomarker classifier based treatment (M-rx) and standard of care based treatment (SOC-rx) are determined before randomization and patients for whom the two treatment strategies agree are not randomized. This design is often much more efficient than that shown in Fig. 1.

domized to either receive Tamoxifen alone or Tamoxifen plus cytotoxic chemotherapy. Randomizing only the patients classified as low risk is much more efficient than randomizing all of the patients. For the case where one treatment reduces the hazard of failure by one-third for the marker positive patients, approximately 200 events are required to obtain 80% statistical power with a two-sided 5% log-rank test for the design of Fig. 2 where only marker positive patients are randomized.

One might argue that treatment determination using a genomic classifier for women with stage I ER positive breast cancer should not be compared to the strategy of giving all such women Tamoxifen plus chemotherapy, because there are practice guidelines available based on tumor size and age that withhold chemotherapy from some patients. Nevertheless, it would still be very inefficient to randomize women to genomic classifier determined therapy or non-genomic practice guidelines determined therapy in which the genomic classifier is measured only on the women randomized to its use. Most of the women will probably receive the same treatment whichever arm they are randomized to. It is much more efficient to perform the assay for measuring the genomic classifier, and then randomize only the women for whom the two treatment strategies differ as indicated in Fig. 2.

The null hypothesis for the design of Fig. 2 is that the marker based treatment selection strategy is equivalent to the standard care treatment selection strategy. In the breast cancer example described above, the marker based treatment selection strategy called for withholding systemic therapy other than tamoxifen for patients predicted to be at low risk of recurrence based on the classifier. The standard of care treatment might incorporate decision making based on established predictive markers. For example, the standard of care might be include treatment with Herceptin for patients whose tumors expressed the Her2/neu receptor. Or the standard of care strategy might involve withholding systemic therapy other than Tamoxifen if the tumor size is below a specified threshold. The design of Fig. 2 requires that the standard of care treatment and the classifier based treatment for each eligible patient be determined before randomization and only those patients for whom the two treatments differ are randomized.

Phase III clinical trials generally attempt to utilize an intervention in a manner that it might be used if adopted in broad clinical practice. For evaluating a diagnostic classifier, a multi-center clinical trial provides the challenges of distributed tissue handling and real time assay performance that would be met in general use. The assays might be performed in multiple laboratories and cannot be batched in time with a single set of reagents as might be done in a retrospective study. Consequently, the prospective clinical trial is the gold standard for external validation of a genomic classifier.

Validation based on a new prospective clinical trial will require a long follow-up time for low risk patients.

In such circumstances it can be useful to conduct a prospectively planned validation using patients treated in a previously conducted prospective multi-center clinical trial if archived tumor specimens are available for the vast majority of patients. The validation study should be prospectively planned with at least as much detail and rigor as for prospective accrual of new patients. Although assaying procedures probably cannot be distributed over time in the same way as for newly accrued patients, assay reproducibility studies should be conducted to demonstrate that the assay has been standardized and quality controlled sufficiently so that such sources of variation are negligible. A written protocol should be developed to ensure that the study is prospectively planned to evaluate the clinical benefit of a completely specified genomic classifier for a defined therapeutic decision in a defined population in a hypothesis testing manner as it would for a prospective clinical trial.

The study of Paik et al. [6] of the OncoType Dx classifier for women with node negative ER positive breast cancer is an example of careful prospective planning of an independent validation study using archived specimens. Their study was based on the observation that although randomizing only the patients classified as low risk is more efficient than randomizing all of the patients, it still would require many patients. It is a therapeutic equivalence trial in the sense that finding no difference in outcome changes clinical practice; consequently it is important to be able to detect small differences. Since the expected recurrence rate is so low, it would take many patients to detect a difference between the treatment arms. But if the recurrence rate is as low as predicted by the classifier, then the benefit of chemotherapy is necessarily extremely small. Consequently, an alternative design for external validation is a single arm study in which the patients classified as low risk are treated with Tamoxifen alone. If, with long follow-up, these patients have a very low recurrence rate, then the classifier is considered validated for providing clinical benefit because it enabled the identification of patients whose prognosis was so good on Tamoxifen monotherapy that they could be spared the toxicity, inconvenience and expense of chemotherapy. This was the approach used by Paik et al. for validation of the OncoType Dx classifier for patients with node negative, estrogen receptor positive breast cancer [6]. The genes that appeared prognostic were initially identified based on published microarray studies. Primers for measuring expression of those genes using RT-PCR of FFPE tissue were developed and a classifier was developed based on archived tissue from NSABP studies. The completely pre-specified classifier was then tested on 668 patients from NSABP B-14 who received tamoxifen alone as systemic therapy. Fifty one percent of the assayed patients fell in the low risk group. They had a distant recurrence rate at 10 years of 6.8 percent (95% confidence interval 4.0 to 9.6). Much higher rates of distant recurrence were seen in the intermediate and high risk groups of the classifier (14.3% and 30.5% respectively).

6. Development of genomic classifiers for experimental drugs

The objective of validation of a genomic classifier differs somewhat for existing therapy compared to an experimental therapy. With existing therapy, the emphasis should be on validation of the clinical benefit of using the classifier. With an experimental therapy, however, the emphasis should be on demonstrating effectiveness of the drug in a population identified by the classifier as being more likely to benefit. Simon and Maitournam [13] demonstrated that use of a genomic classifier for focusing a clinical trial in this manner can result in a dramatic reduction in required sample size, depending on the sensitivity and specificity of the classifier for identifying such patients. Not only can such targeting provide a huge improvement in efficiency in phase III development, it also provides an increased therapeutic ratio of benefit to toxicity and results in a greater proportion of treated patients who benefit.

Simon and Maitournam consider use of the Targeted Design shown in Fig. 3. During pre-clinical and phase I/II clinical development one identifies a fully specified classifier of which patients have a high probability of responding to the experimental drug. That classifier is then used to select patients for phase III trial. This is a form of enrichment design. Table 2 shows the number of events required in order to have 80% statistical power for comparing exponential survival times using the design of Fig. 3 if the treatment results in a halving of the hazard in the patients selected for study using the classifier. The number of events shown in Table 2 is compared to the number of events required in a standard clinical trial if the classifier is not used to select patients for randomization (Table 1). The table assumes that the treatment is not effective for the classifier negative patients. More extensive results on relative efficiency of the targeted and untargeted designs are described by Simon and Maitournam [3,13].

Table 2
Approximate number of events required for 80% power with 5% two-sided log-rank test for comparing arms of design shown in Fig. 3. Only marker + patients are randomized. Treatment hazard ratio for marker + patients is shown in first column. Time-to-event distributions are exponential and all patients are followed to failure

Hazard ratio for marker + patients	Number of events required
0.5	74
0.67	200

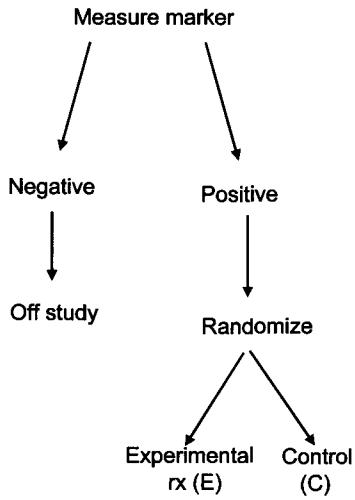

Fig. 3. Targeted clinical trial design for evaluating a new experimental therapy. A biomarker classifier is developed for identifying those patients most likely to respond to the new treatment (E). Only those patients are randomized to E versus the control treatment. The patients predicted less likely to respond (marker negative) are off study. The targeted design is most useful in cases where the biomarker classifier has a strong biological rationale for identifying responsive patients and where it may not be ethically advisable to expose marker negative patients to the new treatment.

Developing a genomic classifier of which patients are likely to benefit for targeting phase III trials may require larger phase II studies. This depends on the type of drug being developed. For example, if the drug is an inhibitor of a kinase mutated in cancer, then there is a natural diagnostic and no genome-wide screening is needed. For many molecularly targeted drugs, however, the appropriate assay for selecting patients is not known and development of a classifier based on comparing expression profiles for phase II responders versus phase II non-responders may be the best approach. In such instances, one may not have sufficient confidence in the genomic classifier developed in phase II to use it for excluding patients in phase III trials as in Fig. 3. It may be better in this case to accept all conventionally eligible patients, and use the classifier in the pre-defined analysis plan.

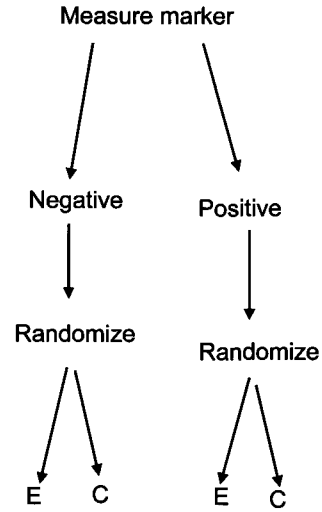

Fig. 4. Stratified analysis design for evaluating a new experimental treatment (E) relative to a control (C). The status of a biomarker based classifier of the likelihood of responding to E is utilized in a prospectively specified analysis plan. The biomarker classifier is not just used for stratifying the randomization. Alternative analysis plans are described in the text.

Figure 4 shows the Marker by Treatment Interaction Design discussed by Sargent et al. [10] and by Pusztai and Hess [7]. Both marker positive and marker negative patients are randomized to the experimental treatment or control. The analysis plan either calls for separate evaluation of the treatment difference in the two marker strata or for testing the hypothesis that the treatment effect is the same in both marker strata. When this design is used for development of an experimental drug, an appropriate analysis plan might be to utilize a preliminary test of interaction; if the interaction is not significant at a pre-specified level, then the experimental treatment is compared to the control overall. If the interaction is significant, then the treatment is compared to the control within the two strata determined by the marker. The sample size planning for such a trial and determination of the appropriate significance level for the preliminary interaction test require further study.

Freidlin and Simon [2] proposed an alternative analysis plan for the design of Fig. 4. They suggested that the overall null hypothesis for all randomized patients is tested at the 0.04 significance level. A portion, e.g. 0.01, of the usual 5 percent false positive rate is reserved for testing the new treatment in the subset predicted by the classifier to be responsive. The analysis starts with a test of the overall null hypothesis, without a preliminary test of interaction. If the overall null hypothesis is rejected, then one concludes that the treatment is effective for the randomized population as

a whole and that the classifier is not needed. If the overall null hypothesis is not rejected at the 0.04 level, then a single subset analysis is conducted; comparing the experimental treatment to the control in the subset of patients predicted by the classifier as being most likely to be responsive to the new treatment. If the null hypothesis is rejected, then the treatment is considered effective for the classifier determined subset. This analysis strategy provides sponsors an incentive for developing genomic classifiers for targeting therapy in a manner that does not unduly deprive them of the possibility of broad labeling indications when justified by the data.

7. Conclusions

Physicians need improved tools for selecting treatments for individual patients. The genomic technologies available today are sufficient to develop such tools. There is not broad understanding of the steps needed to translate research findings of correlations between gene expression and prognosis into robust diagnostics validated to be of clinical utility. This paper has attempted to identify some of the major steps needed for such translation.

Acknowledgements

Thanks to Dr. Wenu Jiang for the computing of Tables 1 and 2.

References

[1] FDA. Draft guidance for industry: Pharmacogenomics data submission. Rockville MD: Food and Drug Administration; 2003.

[2] B. Freidlin and R. Simon, Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients, *Clinical Cancer Research* (2005), in Press.

[3] A. Maitournam and R. Simon, On the efficiency of targeted clinical trials, *Statistics in Medicine* **24** (2005), 329–339.

[4] S. Michiels, S. Koscielny and C. Hill, Prediction of cancer outcome with microarrays: a multiple random validation strategy, *The Lancet* **365**, 488–492.

[5] A.M. Molinaro, R. Simon and R.M. Pfeiffer, Prediction error estimation: A comparison of resampling methods, *Bioinformatics* **21** (2005), 3301–3307.

[6] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin et al., A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, *New England Journal of Medicine* **351** (2004), 2817–2826.

[7] L. Pusztai and K.R. Hess, Clinical trial design for microarray predictive marker discovery and assessment, *Annals of Oncology* **15** (2004), 1731–1737.

[8] M.D. Radmacher, L.M. McShane and R. Simon, A paradigm for class prediction using gene expression profiles, *Journal of Computational Biology* **9** (2002), 505–511.

[9] A. Rosenwald, G. Wright, W.C. Chan, J.M. Connors, E. Campo, R.I. Fisher et al., The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *New England Journal of Medicine* **346** (2002), 1937–1947.

[10] D.J. Sargent, B.A. Conley, C. Allegra and L. Collette, Clinical trial designs for predictive marker validation in cancer treatment trials, *Journal of Clinical Oncology* **23** (2005), 2020–2027.

[11] R. Simon and D.G. Altman, Statistical aspects of prognostic factor studies in oncology, *British Journal of Cancer* **69** (1994), 979–985.

[12] R. Simon and A.P. Lam, *BRB-ArrayTools Users Guide (Version 3.3)*, Bethesda MD: Biometric Research Branch National Cancer Institute; 2005. Report No.: Technical Report 28, http://linus.nci.nih.gob/brb.

[13] R. Simon and A. Maitournam, Evaluating the efficiency of targeted designs for randomized clinical trials, *Clinical Cancer Research* **10** (2004), 6759–6763.

[14] R. Simon, M.D. Radmacher, K. Dobbin and L.M. McShane, Pitfalls in the analysis of DNA microarray data: Class prediction methods, *Journal of the National Cancer Institute* **95** (2003), 14–18.

[15] R.M. Simon, E.L. Korn, L.M. McShane, M.D. Radmacher, G.W. Wright and Y. Zhao, *Design and analysis of DNA microarray investigations*, New York: Springer; 2003.