Cancer Therapy: Clinical

Clinical Cancer Research

The Cross-Validated Adaptive Signature Design

Boris Freidlin[1], Wenyu Jiang[2], and Richard Simon[1]

Abstract

Purpose: Many anticancer therapies benefit only a subset of treated patients and may be overlooked by the traditional broad eligibility approach to design phase III clinical trials. New biotechnologies such as microarrays can be used to identify the patients that are most likely to benefit from anticancer therapies. However, due to the high-dimensional nature of the genomic data, developing a reliable classifier by the time the definitive phase III trail is designed may not be feasible.

Experimental Design: Previously, Freidlin and Simon (*Clinical Cancer Research*, 2005) introduced the adaptive signature design that combines a prospective development of a sensitive patient classifier and a properly powered test for overall effect in a single pivotal trial. In this article, we propose a cross-validation extension of the adaptive signature design that optimizes the efficiency of both the classifier development and the validation components of the design.

Results: The new design is evaluated through simulations and is applied to data from a randomized breast cancer trial.

Conclusion: The cross-validation approach is shown to considerably improve the performance of the adaptive signature design. We also describe approaches to the estimation of the treatment effect for the identified sensitive subpopulation. *Clin Cancer Res; 16(2); 691–8. ©2010 AACR.*

Due to the molecular heterogeneity of most human cancers, only a subset of treated patients benefit from a given therapy. This is particularly relevant for the new generation of anticancer agents that target specific molecular pathways (1–3). Genomic (or proteinomic) technologies such as microarrays provide powerful tools for identifying a genetic signature (diagnostic test) for patients who are most likely to benefit from a targeted agent. Ideally, such diagnostic test should be developed and validated before commencing the definitive phase III trial (4). However, due to the complexity of signaling pathways and the large number of genes available for analysis, the development of a reliable diagnostic classifier using early nonrandomized phase II data is often not feasible. Conducting a phase III randomized clinical trial (RCT) requires considerable time and resources. Therefore, clinical trial designs that allow combining the definitive evaluation of a new agent with the development of the companion diagnostic test can considerably speed up the introduction of new cancer therapies.

Previously, the adaptive signature design (ASD) has been proposed for settings where a signature to identify sensitive patients is not available (5). The design combines the prospective development of a pharmacogenomic diagnostic test (signature) to select sensitive patients with a properly powered test for overall effect. It was shown that when the proportion of patients sensitive to the new drug is low, the ASD substantially reduces the chance of false rejection of effective new treatments. When the new treatment is broadly effective, the power of the adaptive design to detect the overall effect is similar to that of the traditional design.

The signature component of the ASD carries out signature development and validation on the mutually exclusive subgroups of patients (e.g., half of the study population is used to develop a signature and another half to validate it). Although the conceptual simplicity of this approach is appealing, it also limits its power as only half of the patients are used for signature development and half for validation. This is especially relevant in the present setting because (*a*) signature development in high dimensional data requires large sample sizes, and (*b*) when the fraction of sensitive patients is low, a large number of patients needs to be screened to identify the sufficient number of sensitive patients to achieve acceptable power.

In this article, we describe an extension of the ASD in which signature development and validation are embedded in a complete cross-validation procedure. This allows the use of virtually the entire study population in both signature development and validation steps. We develop a procedure that preserves the study-wise type I error while substantially increasing the statistical power for establishing a statistically significant treatment effect for an identified subset of patients who benefit from the experimental treatment. We also examine approaches to estimation of treatment effect for the identified sensitive subset.

Authors' Affiliations: [1]Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland and [2]Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada

Corresponding Author: Boris Freidlin, Biometric Research Branch, EPN-8122, National Cancer Institute, Bethesda, MD 20892. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov.

doi: 10.1158/1078-0432.CCR-09-1357

Translational Relevance

Traditional broad eligibility approaches to design definitive randomized clinical trials may overlook new targeted anticancer therapies that benefit only a subset of treated patients. The proposed clinical trial design provides an effective and valid approach to combining the prospective development of a classifier for patients who are most likely to benefit from a targeted agent with a properly powered test for overall treatment effect in a single phase III clinical trial. This can help to dramatically accelerate the development of new anticancer therapies.

Materials and Methods

The final stage of clinical drug development usually requires showing that the addition of the new agent to the standard treatment is beneficial in an RCT that randomly assigns patients to the combination of the new and standard treatment (arm E) or the standard treatment alone (arm C).

The following presentation is expressed in terms of a predictive classifier based on DNA microarray gene expression signatures, but the design is easily adapted to classifier development based on single nucleotide polymorphism genotyping, proteomic profiling, or even selection from single gene or protein candidate classifiers. For the development based on gene expression data, we have used the following modeling assumptions: among H evaluated genes, there is a subset of L predictive ("sensitivity") genes. The identities of the predictive genes are unknown, but responsiveness to treatment is influenced by the predictive genes through the following model. The model is based on a binary response end point as in the ASD but is easily generalized to a proportional hazards model for survival or disease-free survival model. For a given patient, let p denote the probability of response, t denote the treatment indicator ($t = 0$ for arm C and $t = 1$ for arm E), and $x_1,\ldots,x_L$ denote the levels of expression for the L unknown predictive genes. Then

$$\log\left(\frac{p}{1-p}\right) = \mu + \lambda t + \xi_1 x_1 + \ldots \xi_L x_L + \gamma_1 t x_1 + \ldots + \gamma_L t x_L \quad (1)$$

where λ is the treatment main effect that all patients experience regardless of their gene expression levels, ξ_i is main effect for the i^{th} sensitivity gene, and γ_i is the treatment-expression interaction effect that reflects the degree by which the difference in treatment arms is influenced by the i^{th} sensitivity gene expression level. To simplify the presentation main effects and the treatment expression, interactions for the nonpredictive genes are assumed to be 0.

If the interaction parameters are positive, patients who overexpress the predictive genes have a higher probability of response on arm E compared with arm C. We assume that a fraction of the patient population overexpresses some (but not necessarily all) of the predictive genes. These patients are called "sensitive."

First, we briefly describe the original ASD. Then, the cross-validated version of the design (CVASD) is introduced.

ASD design. Consider designing a phase III trial with an overall type I error α. Similar to the traditional RCT design, the ASD is based on randomizing a total of N patients between the two treatment arms. At the completion of the trial, the final analysis begins with an overall comparison between the new and the standard treatment arms using the data from all N patients. If the comparison is significant at a prespecified significance level α_1 ($\alpha_1 < \alpha$), then the new treatment is considered beneficial for the broad population of patients. Otherwise, the design proceeds to the signature development/validation stage. The study patients are divided into two cohorts: (*a*) the signature validation cohort (often called the test or validation data set) that contains M patients and (*b*) the signature development cohort (often called the training data set) that contains the remaining N-M patients. Using the development cohort patients only, a signature, which tries to identify patients that have better outcomes on the new therapy than the standard therapy, is developed. The signature is then applied to the validation cohort, and "signature-positive" patients called "sensitive patients" are identified. Outcomes for patients in the sensitive subset of the validation cohort who received the new therapy are compared with the outcomes for patients in the sensitive subset of the validation cohort who received the standard therapy. A statistical significance test is done at the significance threshold $\alpha_2 = \alpha - \alpha_1$ to ensure that overall type I error of the design is no greater than $\alpha = \alpha_1 + \alpha_2$. The allocation of the experiment-wise significance level α between the overall and subset tests in a particular implementation of the design should be based on the strength of the existing evidence for the distribution of the new drug activity in study population. To preserve the ability of the procedure to detect an overall effect, we recommend setting α_1 in the 50% to 80% of the α range. For example, with 80% to 20% allocation setting, $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$ corresponds to a procedure-wise α level of 0.05.

A large variety of algorithms for developing a classifier based on patients accrued during stage 1 could be envisioned. To illustrate the adaptive designs, we use an approach based on machine learning voting methods (6).

Step 1: Using data from stage 1 patients, for each gene j, fit the single gene logistic model $\text{logit}(p) = \mu + \lambda_j t + \beta_j t x_j$. Note the genes that have treatment-expression interaction ($\hat{\beta}_j$) that is significant at a predetermined level η.

Step 2: Classify stage 2 patients as sensitive or nonsensitive to the new treatment based on the genes with significant interactions in Step 1. Patient in stage 2 is designated sensitive if the predicted new versus control arm odds ratio exceeds a specified threshold (R)

for at least G of the significant genes j (that is, $e^{\hat{\lambda}_j + \hat{\beta}_j x_j} > R$).

A list Θ of plausible sets of the three tuning parameters (η, R, G) should be identified prospectively based on the range of feasible models. Freidlin and Simon (5) described the selection of tuning parameter set from the list.

CVASD design. Because the signature development and the subset effect testing stages of the ASD are done on two nonoverlapping subpopulations, the nominal significance level of the entire signature development testing procedure is preserved at the nominal α_2 level. However, this conceptual simplicity results in the loss of efficiency (because only a portion of the trial patients contributes to each stage). To allow a more efficient use of all available data, the CVASD uses cross-validation approach for the signature development and subset effect testing. A K-fold cross-validation for an N-patient trial proceeds as follows: first, the trial population is split into an M-patient validation cohort and an (N–M)-patient development cohort (where $M = N/K$). Let D_k denote the set of patients in the k^{th} development cohort, and V_k as the set of patients in the corresponding validation cohort. For each D_k ($k = 1,\ldots,K$), a predictive signature is developed. The signature is applied to identify a sensitive patient subset S_k of V_k. This procedure is repeated K times over all M-patient–nonoverlapping validation cohorts V_k (and the corresponding D_k). Each study patient appears in exactly one of the validation cohorts.

At the end of the cross-validation procedure, each of the study patients is classified as either sensitive or not. The sensitive patient subset of the entire study population is identified as $S = \cup_{k=1}^{K} S_k$. The outcomes for the sensitive patients who received the experimental therapy can be compared with the outcomes for the sensitive patients who received the standard therapy. Because this subset of sensitive patients is obtained by cross-validation, the standard asymptotic theory does not apply. We use a permutation method (7) to obtain a valid P value for a given test statistic T (where T is testing presence of treatment effect in the sensitive patient subset). The permutation distribution can be approximated by a resampling-based approach. First, statistic T is evaluated on the observed data. Then, a permuted data set is constructed by randomly permuting treatment labels. The entire cross-validation procedure is repeated for the permuted data set and the corresponding test statistic T^* is computed. This procedure is repeated for B permuted data sets. The permutation P value is given by

$$\frac{1 + \textit{number of permutations where } T^* \geq T}{1 + B}.$$

It is important for all steps of the signature development algorithm (including the selection of tuning parameters) to be incorporated into each loop of the cross-validation procedure for each permuted data set to obtain a valid P value (7, 8). In CVASD, this is accomplished as follows: a list Θ of plausible tuning parameter sets (η, R, G) is prespecified; then on each (N-M)-patient development cohort D_k, a nested inner loop of K-fold cross-validation is

Table 1. Empirical power: ASD and CVASD as a function of the subset treatment effect, 10% of patients are sensitive

Test	ASD	CVASD (10-fold cross-validation)
Ninety percent response in sensitive patients on the experimental arm and 25% in all other patients		
Overall 0.05 level test	0.301	0.291
Overall 0.04 level test	0.261	0.256
Sensitive subset 0.01 level test	0.495	0.880
Overall power	0.605	0.909
Eighty percent response in sensitive patients on the experimental arm and 25% in all other patients		
Overall 0.05 level test	0.232	0.240
Overall 0.04 level test	0.203	0.209
Sensitive subset 0.01 level test	0.202	0.661
Overall power	0.348	0.714
Seventy percent response in sensitive patients on the experimental arm and 25% in all other patients		
Overall 0.05 level test	0.177	0.183
Overall 0.04 level test	0.147	0.155
Sensitive subset 0.01 level test	0.066	0.371
Overall power	0.191	0.450
Sixty percent response in sensitive patients on the experimental arm and 25% in all other patients		
Overall 0.05 level test	0.121	0.129
Overall 0.04 level test	0.105	0.107
Sensitive subset 0.01 level test	0.015	0.135
Overall power	0.115	0.229

Table 2. Empirical power: ASD and CVASD as a function of the fraction of sensitive patients, 70% response in sensitive patients on the experimental arm and 25% in all other patients

Test	ASD	CVASD (10-fold cross-validation)
Twenty percent of patients are sensitive		
Overall 0.05 level test	0.518	0.503
Overall 0.04 level test	0.478	0.471
Sensitive subset 0.01 level test	0.192	0.588
Overall power	0.543	0.731
Thirty percent of patients are sensitive		
Overall 0.05 level test	0.846	0.838
Overall 0.04 level test	0.822	0.808
Sensitive subset 0.01 level test	0.295	0.723
Overall power	0.836	0.918
Forty percent of patients are sensitive		
Overall 0.05 level test	0.969	0.961
Overall 0.04 level test	0.962	0.955
Sensitive subset 0.01 level test	0.412	0.812
Overall power	0.963	0.972

applied to obtain a set of sensitive patients corresponding to each of the parameter sets. The tuning parameter set corresponding to the sensitive subset with smallest P value (for the difference between arms) is selected for use in the corresponding development cohort. Finally, to ensure the strict validity and reproducibility of the procedure, patient allocation to the cross-validation cohorts D_k and V_k ($k = 1, \ldots, K$) should be prospectively defined.

Results

We conducted a simulation study to evaluate the performance of ASD and CVASD with $K = 10$ (10-fold cross-validation). We assumed a 10,000-gene array with $L = 10$ predictive genes. A previous simulation study indicated that the design performance is similar over a range of values of L (5). Gene main effects, ξ_i, were assumed to be 0. Treatment-expression interaction levels were kept constant across predictive genes ($\gamma = \gamma_1 = \gamma_2 = \ldots = \gamma_L$). An intercept ($\mu$) value corresponding to a control arm response rate of 25% was used (previous simulation suggested similar results for other intercept values; ref. 5). The simulations were based on 1,000 replications and $B = 99$ (see the Appendix A for more details).

The simulations are reported in Tables 1 to 6 and represent a clinical trial in which $N = 400$ patients were randomized between the experimental and control arms. Both CVASD and ASD procedures use 80% to 20% α-level split with an overall 0.05 (two sided) significance level (corresponding to $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$). We tabulated the empirical powers of the overall arm comparison at 0.05 and 0.04 significance levels, and of the comparison in the selected sensitive subset at 0.01 two-sided significance levels. In addition, the overall empirical power of the adaptive designs is calculated as the percentage of replications with either positive overall 0.04 level test or positive 0.01 level sensitive subset test. Because ASD and CVASD simulations were run independently, we report the empirical power for the overall tests for each procedure separately (note that the small discrepancies between the power of the overall 0.05 or 0.04 level tests for the two procedures reflect the random variation in the simulation results).

We start by considering a situation where the benefit of the new drug is restricted to a small fraction of patients: 10% of the eligible patient population that overexpresses the predictive genes. Table 1 presents the results for a range of treatment effects. The first panel corresponds to a very strong subset effect: 90% response rate in sensitive patients on the experimental arm and a 25% response rate in all other patients (nonsensitive patients on the experimental arm and all control arm patients). In both ASD and CVASD, the overall difference was declared with a 26% probability (using a 0.04 level overall test). In ASD, the sensitive subset test was significant (at 0.01 level) in 50% of cases resulting in a 61% overall power for the design (a 61% probability of either detecting a significant overall effect or a significant subset effect). For CVASD, the sensitive subset test was significant (at 0.01 level) in 88% of cases resulting in a 91% overall power (as a reference, note that in this case, the traditional broad eligibility design that relies on an overall 0.05 level test has a 30% power to detect difference between the arms). CVASD shows a better ability to detect subset effect relative to

Table 3. Empirical power: ASD and CVASD as a function of the fraction of sensitive patients, 60% response in sensitive patients on the experimental arm and 25% in all other patients

Test	ASD	CVASD (10-fold cross-validation)
Twenty percent of patients are sensitive		
Overall 0.05 level test	0.359	0.349
Overall 0.04 level test	0.323	0.313
Sensitive subset 0.01 level test	0.051	0.196
Overall power	0.342	0.421
Thirty percent of patients are sensitive		
Overall 0.05 level test	0.620	0.622
Overall 0.04 level test	0.584	0.582
Sensitive subset 0.01 level test	0.047	0.254
Overall power	0.589	0.641
Forty percent of patients are sensitive		
Overall 0.05 level test	0.861	0.856
Overall 0.04 level test	0.841	0.829
Sensitive subset 0.01 level test	0.089	0.274
Overall power	0.842	0.843

Table 4. Empirical power: no subset effect

Test	ASD	CVASD (10-fold cross-validation)
No subset effect: 35% response in all patients on the experimental arm, 25% response rate in all patients on the control arm		
Overall 0.05 level test	0.572	0.594
Overall 0.04 level test	0.534	0.554
Sensitive subset 0.01 level test	0.009	0*
Overall power	0.534	0.554
No treatment effect: 25% response in all patients		
Overall 0.05 level test	0.050	0.056
Overall 0.04 level test	0.038	0.048
Sensitive subset 0.01 level test	0.001	0*
Overall power	0.038	0.048

*When no sensitive patients are identified, the test statistic is given the lowest possible value.

ASD across the range of treatment effects (Table 1, *panels 2-4*). In fact, the relative gain in power for CVASD versus ASD increases as the magnitude of the treatment effect is reduced. For example, when the response rate in sensitive patients on the experiment arm is 70%, the sensitive subset test was significant in 37% of cases in CVASD versus 7% in ASD (45% versus 19% overall power). Moreover, for smaller treatment effects (*panels 3 and 4*) the power advantage for the adaptive designs versus the traditional broad eligibility approach (that relies on overall effect test) is limited to CVASD.

Tables 2 and 3 illustrate ASD and CVASD performance with increasing fraction of the sensitive patients (strong subset effect in Table 2 and moderate subset effect in Table 3). Across the board, CVASD provides an improved detection of the subset effect. As the sensitive patient fraction increases, so does the ability of an overall test to detect treatment effect in the entire study population. Thus, in terms of the overall power, the difference between ASD, CVASD, and the traditional design becomes smaller for higher fractions of sensitive patients, and the overall power is similar when sensitive fraction is 40% or above. For example, when 40% of the study patients are sensitive, overall powers of the ASD, CVASD, and the traditional design are 84%, 84%, and 86%, respectively (Table 3, *panel 3*).

When all patients benefit equally from the new treatment, both CVASD and ASD correctly indicate the absence of sensitive subpopulation (Table 4, *panel 1*). At the same time, the adaptive procedures preserve the power for detection of the overall effect. When there is no benefit from the new therapy (overall or in a subset), the adaptive procedures preserve the overall type I error rate (Table 4, *panel 2*).

An important issue in the application of an adaptive design is the interpretation of a positive study. In particular, it is important to characterize the sensitive subset and to have an estimate of the treatment effect in the sensitive subpopulation. Consider a study that used CVASD and obtained a significant treatment effect in a subpopulation but not for the overall population. The final classifier to identify patients that benefit from the new drug is obtained by applying the signature development algorithm to the entire study population. A relevant measure of the treatment effect is then the predicted treatment effect in the future patients that are classified as sensitive by the final signature. Two possible estimators for the predicted effect are (*a*) the empirical treatment effect observed in patients identified as sensitive by applying the final signature to the entire study population (we refer to this as the resubstitution estimator because it is obtained by applying the signature to the patients that were used to develop the signature), and (*b*) the treatment effect observed in patients identified as sensitive by *K*-fold cross-validation using the tuning parameters selected from the entire study population (as in the resubstitution estimator above; we refer to this as CV estimator). The performance of the two estimators is illustrated in Table 5 for two settings with 10% fraction of sensitive patients. The table is based on calculating both the resubstitution and the 10-fold CV treatment effect estimates on a simulated 400-patient trial. The treatment effect observed by applying the final signature to an independent set of 1,000 patients is used as a reference value for the true predicted effect ("Predicted rates" column in Table 5). The results were then averaged over 1,000 simulations. As expected, the resubstitution estimate tends to overestimate the predicted effect because it uses the same data to develop the signature and estimate the treatment effect. On the other hand, the 10-fold CV estimate tends to underestimate the predicted effect. The conservatism presumably results from the suboptimality of the classifiers developed and optimized based on nested

Table 5. Resubstitution and CV estimates of the predicted treatment effect in the sensitive subpopulation (10% of the patients are sensitive)

	Predicted rates	Resubstitution rates	CV rates
Ninety percent response in sensitive patients, 25% response in all other patients			
Experimental arm	90%	91%	87%
Control arm	25%	23%	27%
Difference	65%	68%	60%
Seventy percent response in sensitive patients, 25% response in all other patients			
Experimental arm	74%	88%	65%
Control arm	25%	19%	28%
difference	49%	69%	37%

Table 6. Results of applying CVASD to Bonnefoi et al. (2007) EORTC 10994 Neoadjuvant breast cancer data

Overall comparison

***P* = 0.79**

Arm	Observed pCR rate (%) (no of patients)
FEC	42% (66)
TET	45% (58)

Sensitive subset comparison

P* = 0.006

Arm	Estimates of pCR rates in the sensitive subpopulation	
	Resubstitution	CV
FEC	20% (15)	29% (14)
TET	100% (8)	83% (12)

**P* value based on permutation distribution of the cross-validated treatment effect in sensitive subset.

cross-validation procedure compared with the classifier developed using the full data set.

To illustrate our approach, we searched the National Center for Biotechnology Information Gene Expression Omnibus depository for publicly available gene expression data from a RCT. The only randomized cancer trial data with both expression and clinical outcome available that we were able to identify were data on a subset of 124 hormone receptor–negative breast cancer patients treated on EORTC 10994 (reported in ref. 9 and available at National Center for Biotechnology Information Web site).[3] EORTC 10994 was a phase III neoadjuvant breast cancer RCT that compared nontaxane regimen of 5-fluorouracil, cyclophosphamide, and epirubicin (FEC) with a taxane regimen of epirubicin and docetaxel (TET). In 66 patients treated with FEC, 28 had pathologic complete response (pCR), and in 58 patients treated with TET, 26 had pCR. CVASD was applied to these data (see Appendix B for details), and results are presented in Table 6. There was no overall difference in pCR rates between TET and FEC arms (pCR rates 45% and 42%, respectively; $P = 0.79$). The CVASD algorithm indicated the existence of a significant ($P = 0.006$) sensitive subset where TET is substantially more effective than FEC: the conservative (CV) estimate of the treatment effect was 83% pCR (TET) versus 29% pCR (FEC). Although providing a biological rationale for the sensitive patient signature is beyond the scope of this article, we note that two of the probes in the signature (Hs.310359.0.A1_3p_at and g4507484_3p_a_at) are related to the mitogen-activated protein kinase pathway that has been reported to be associated with anthracycline resistance in hormone receptor–negative breast cancer (10).

[3] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6861

As this is a retrospective application of CVASD to a subset of a reported RCT, these results will need an independent confirmation.

Discussion

Our results show that the cross-validation approach can considerably enhance the ASD performance. Cross-validation permits the maximization of the portion of study patients contributing to the development of the diagnostic signature [as shown by Molinaro et al. (11), this is critical in the high-dimensional data setting where the sample size or signal to noise ratio is limited]. Cross-validation also maximizes the size of the sensitive patient subset used to test (validate) the signature (this is important in settings where the fraction of the sensitive patients is small).

In this presentation, we used 80% to 20% allocation of the error rates between the overall and subset tests. This allocation represents a conservative approach that is aimed at preserving the ability of detecting the overall treatment effect without increasing the overall sample size. Depending on the amount of preliminary evidence that the treatment effect is limited to a subpopulation, one might allocate a higher proportion (up to 50%) of the overall error to the subset effect (i.e., using $\alpha_1 = 0.025$ and $\alpha_2 = 0.025$ for an overall 0.05 level design). The study sample size could be increased to achieve a desired power for the overall analysis or for the subset analysis. Sample size depends not only on the proportion of patients in the sensitive subset and the treatment effect in that subset, but also on aspects of the data used for classifier development. We plan to study sample size planning for the CVASD.

An important step in interpreting a trial that indicates that the effect of the new therapy is limited to a subset of patients is to provide an explicitly defined diagnostic test to identify the subpopulation of the future patients

that are most likely to benefit and to quantify the beneficial effect in that subpopulation. In the CVASD design, the final diagnostic test is obtained by applying the signature development procedure to the entire study population; a conservative estimate of the treatment effect in the corresponding subpopulation of the future patients can be obtained by a 10-fold CV method. Although the conservatism of the estimate is preferable to anticonservative of the resubstitution estimate, in the future, we plan to investigate the parameters influencing the accuracy of the estimator of treatment effect.

Our approach uses cross-validation for two purposes. First, to define a valid test of the null hypothesis that there is no sensitive subgroup of patients that benefit from the new regimen compared with control. The second purpose of the cross-validation is to provide a conservative estimate of the treatment effect that would be seen in the application of the classifier defined based on the full data set to identify sensitive patients in the future. This estimate is obtained by conservatively estimating the mean treatment effect in the sensitive subgroup in the reapplication of the algorithm to resampling the study population. This is analogous to the use of cross-validation or bootstrap methods for evaluating the predictive accuracy of simple binary classifiers (12). Molinaro et al. (11) have shown that CV can provide similar level of validity while being more efficient than the split-sample approach (the approach used in the original ASD design). As CV and split-sample are methods for internal validation, however, there is value in external validation for both methods.

Reselection of the informative genes for different loops of the cross-validation is essential to the validity of the approach (13). The fact that the selected gene set may not be stable when portions of the cases are omitted does not mean that the classifications are unstable or that the classifier developed on the full data set will not predict accurately for independent data. Good genomic signatures are generally not unique. Due to the correlation in gene expression levels, different data sets may result in seemingly different and even nonoverlapping signatures with good predictive characteristics (14). We have used the method of gene selection and model building previously developed for the original ASD. Genes are selected based on their differential effect on outcome between the two treatment groups (i.e., interaction). A weighted voting classifier based on the selected genes is used to classify individual patients as sensitive to the new regimen compared with control. Although this approach is based on methods widely used in the machine learning literature, many other approaches are possible. The current literature on classifier development in bioinformatics is dominated by methods for binary classification that is not directly applicable to the RCT setting. In many cases, classification accuracy is more determined by the difficulty of the problem than by the gene selection method or classifier type used. This is not always the case however, and certainly, further research in classifiers for use in the analysis of RCTs with high dimensional data is warranted.

Appendix A: Simulation Details

In the simulation study, gene expression levels were generated as follows:

(1) For predictive genes in sensitive patients: using a multivariate normal distribution with mean 1, variance $\sigma_1^2 = 0.25$, and correlation $\rho = 0$,
(2) For predictive genes in nonsensitive patients: using multivariate normal distribution with mean 0, variance $\sigma_2^2 = 0.01$, and correlation $\rho = 0$,
(3) For nonpredictive genes: using multivariate normal with mean 0, variance $\sigma_0^2 = 0.25$, and correlation $\rho = 0$ in all patients.

We used $\rho = 0$ as previous simulations showed similar results with correlated data (5).

To optimize the computation time for the simulation, instead of fitting individual logistic regression in step 1 of the signature development algorithm, we used the standardize difference between average gene expressions in responders versus nonresponders in two arms:

$$[(\bar{x}_{R,E} - \bar{x}_{NR,E}) - (\bar{x}_{R,C} - \bar{x}_{NR,C})] \Big/ \sqrt{\sigma^2(1/n_{R,E} + 1/n_{NR,E} + 1/n_{R,C} + 1/n_{NR,C})}$$

Where
$\bar{x}_{R,E}(n_{R,E})$, $\bar{x}_{NR,E}(n_{NR,E})$ $\bar{x}_{R,C}(n_{R,C})$, and $\bar{x}_{NR,C}(n_{NR,C})$

are mean expression (sample size) for responders on arm E, nonresponders on arm E, responders on arm C, and nonresponders on arm C, respectively.

In CVASD, to further limit the simulation time, in the selection of the tuning parameters in each permutation run, only the first cross-validation subset, D_1, was used to select tuning parameters that were used throughout that run. The following list of plausible tuning parameters (η, R, G) was used; {(0.02,10,4), (0.02, 12, 3), (0.02, 20, 1)}.

Appendix B: Application of CVASD to Bonnefoi et al. 2007 EORTC 10994 Data

The gene expression data for 124 patients were obtained on the Affymetrix X3P microarray with 61,000 probe sets (note that Bonnefoi et al. reported results on 125 patients but gene expression data on National Center for Biotechnology Information Gene Expression Omnibus Web site was only available on 124 patients). To adjust for the relatively small sample size, the expression data were filtered by restricting the analysis to the 5,000 probes with the highest variability. We then applied CVASD using the following list of tuning parameter sets {(0.02,10,4), (0.02, 12, 3), (0.02, 20, 1)}. Sensitive subset P value is based on 999 permutations.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 5/28/09; revised 10/5/09; accepted 10/27/09; published OnlineFirst 1/12/10.

References

1. Balis FM. Evolution of anticancer drug discovery and the role of cell based screening. J Natl Cancer Inst 2002;94:78–9.
2. Schilsky RL. End points in cancer clinical trials and the drug approval process. Clin Cancer Res 2002;8:935–8.
3. Rothenberg ML, Carbone DP, Johnson DH. Improving the evaluation of new cancer treatments: challenges and opportunities. Nat Rev Cancer 2003;3:303–9.
4. Hoering A, LeBlanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. Clin Cancer Res 2008;14:4358–67.
5. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 2005;11:7872–8.
6. Breiman L. Bagging predictors. Machine Learn 1996;24:123–40.
7. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and Analysis of DNA Microarray Investigations. New-York: Springer; 2004.
8. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 2006;7:91.
9. Bonnefoi H, Potti A, Delorenzi M, et al. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. Lancet Oncol 2007;8:1071–8.
10. Derin D, Eralp Y, Ozluk Y, et al. Lower level of MAPK expression is associated with anthracycline resistance and decreased survival in patients with hormone receptor negative breast cancer. Cancer Invest 2008;26:671–9.
11. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21:3301–7.
12. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall; 1993.
13. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
14. Fan C, Oh DS, Wessels L, et al. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006;355:560–9.