Oncology Clinical Trials in the Genomic Era

Richard Simon and Jyothi Subramanian

Abstract Developments in genomics are providing a biological basis for the heterogeneity of clinical course and response to treatment that have long been apparent to clinicians. The ability to molecularly characterize human diseases presents new opportunities to develop more effective treatments and new challenges for the design and analysis of clinical trials.

In oncology, treatment of broad populations with regimens that benefit a minority of patients is less economically sustainable with expensive molecularly targeted therapeutics. The established molecular heterogeneity of human diseases requires the development of new paradigms for the design and analysis of randomized clinical trials as a reliable basis for predictive medicine.

We review prospective designs for the development of new therapeutics and predictive biomarkers to inform their use. We cover designs for a wide range of settings. At one extreme is the development of a new drug with a single candidate biomarker and strong biological evidence that marker negative patients are unlikely to benefit from the new drug. At the other extreme are phase III clinical trials involving both genome-wide discovery of a predictive classifier and internal validation of that classifier. We have outlined a prediction-based approach to the analysis of randomized clinical trials that both preserves the type I error and provides a reliable internally validated basis for predicting which patients are most likely or unlikely to benefit from a new regimen.

R. Simon (✉)
Biometric Research Branch, National Cancer Institute, Emmes Corporation, 9000 Rockville Pike, MSC7434, Bethesda, Rockville MD 20892-7434, USA
e-mail: rsimon@nih.gov

T.R. Fleming and B.S. Weir (eds.), *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, Lecture Notes in Statistics 1205,
DOI 10.1007/978-1-4614-5245-4_5,

1 Introduction

This dominant paradigm for major clinical trials today involves using broad eligibility criteria and to randomly assign an experimental treatment or control to test a single null hypothesis that a single clinical outcome measure is on average unimproved by the experimental treatment. Although it is recognized that no two patients are identical, it is implicitly assumed that all have the same disease and that treatment benefit, if it exists, differs only in magnitude among subsets of patients. In this paradigm, subset analysis is viewed with suspicion and is considered only exploratory for purpose of hypothesis generation for future studies. All aspects of multiplicity are accounted for in the test of a single primary null hypothesis. Large sample sizes and multicenter participation are the rule in order to be able to detect small average absolute treatment effects.

The emphasis on broad eligibility criteria has been based on a concern that drugs found effective in clinical trials might subsequently be used in broader patient populations [1, 2]. Some clinical trials even abandoned formal eligibility criteria in favor of the "uncertainty principle" which stated that if the individual physician was uncertain about which treatment might be better for a patient, then that patient was eligible [3]. The focus on ignoring subset analysis unless the overall null hypothesis can be rejected is based on concern about data dredging, the assumption that qualitative interactions are unlikely [3, 4] and that drugs are inexpensive and without serious side effects. For oncology today, none of those assumptions are appropriate. Treating the majority for the benefit of the minority is no longer an effective public health strategy.

Randomized clinical trials have made important contributions to modern medicine and public health, but they have also led to the overtreatment of broad populations of patients, most of whom don't benefit from the increasingly expensive drugs and procedures shown to have statistically significant average treatment effects in increasingly large clinical trials. Fortunately the tools of biotechnology and genomics are providing the tools to identify the subsets of patients who benefit from treatments.

Developments in our understanding of the genomic basis of cancer have indicated that cancers of most primary sites (e.g., lung and breast) represent a heterogeneous collection of diseases that differ in pathophysiology, natural history, and sensitivity to treatment. Recent results have demonstrated that these diseases differ with regard to the mutations that cause them and drive their invasion. The new understanding of heterogeneous nature of tumors of the same primary site leads to new challenges with regard to clinical trial design. Today we are challenged to develop a new paradigm of clinical trial design and analysis that enables development of a predictive medicine that is science based and reliable. Physicians have always known that cancers of the same primary site were heterogeneous with regard to natural history and response to treatment. This understanding led to conflicts with statisticians over the use of subset analysis in the analysis of clinical trials. Although most statisticians expressed little interest in subset analysis methods [5], many

practitioners rejected the results of clinical trials whose conclusions were based on average effects. Today we have powerful tools for characterizing the tumors biologically and using this characterization as a basis for the design and analysis of clinical trials.

Most oncology drugs are being developed for defined molecular targets but the traditional diagnostic classification schemes that are the basis for clinical trial eligibility criteria include patients whose tumors are and are not driven by deregulation of those targets. For many drugs, the targets are well understood and there is a compelling biological basis for restricting development to the subset of patients whose tumors are characterized by deregulation of the drug target. For other drugs there is more uncertainty about the target, and how to measure whether the target is driving tumor invasion in an individual patient [6]. It is clear that the primary analysis of the new generation of oncology clinical trials must consist of more than just treating the traditionally broad patient populations and testing the null hypothesis of no average effect. But it is also clear that the tradition of post-hoc data dredging subset analysis is not an adequate basis for predictive oncology. We need prospective analysis plans that provide for both preservation of the type I experiment-wise error rate and for focused predictive analyses that can be used to reliably select patients in clinical practice for use of the new regimen [7]. These two primary objectives are not inconsistent, and clinical trials should be sized for both purposes.

The following sections summarize some of the designs that have been developed for the new generation of cancer clinical trials. Developing new treatments with companion diagnostics or predictive biomarkers for identifying the patients who benefit does not make drug development simpler, quicker, or cheaper as is sometimes claimed. Actually it makes drug development more complex and probably more expensive. But for many new oncology drugs it is the only science-based approach and should increase the chance of success. It may also lead to more consistency in results among trials and has obvious benefits for reducing the number of patients who ultimately receive expensive drugs which expose them risks of adverse events but no benefit. This approach also has great potential value for controlling societal expenditures on health care.

The ideal approach is prospective drug development with a companion diagnostic [7]. This approach, which is being used extensively today in oncology involves (1) Development of a completely specified predictive classifier using preclinical and early phase clinical studies. The classifier may be based on a single gene or protein or a composite score incorporating the levels of expression of multiple genes. (2) Development of an analytically validated test for measurement of that classifier. Analytically validated means that the test accurately measures what it is supposed to measure, or if there is no gold-standard measurement, that the test is reproducible and robust. (3) Use of that completely specified classifier and analytically validated test to design and analyze a new clinical trial to evaluate the effectiveness of that drug and how the effectiveness relates to the classifier. The guiding principle is that the data used to develop the classifier should be distinct from the Phase III data used to test hypotheses about treatment effects in subsets determined by the classifier. This is in contrast to the typical paradigm in which multiple variables are measured

using non-analytically validated tests and then performing an exploratory analysis that requires confirmation in a subsequent study. In the enrichment and stratified designs described below, biomarker discovery is performed prior to the phase III trial and a single completely specified classifier is used in the trial. We will also discuss designs and prospective analysis plans that incorporate multiple candidate classifiers or even broader classifier development and evaluation in the same clinical trial. But in all of these designs, the analysis plans are carefully pre-specified to ensure that treatment effects in classifier-based subsets are unbiasedly estimated and that overall type I error is preserved.

2 Targeted (Enrichment) Designs

Designs in which the eligibility criteria restrict the clinical trial to those patients considered most likely to benefit from the experimental drug are called "targeted designs" or "enrichment designs." With an enrichment design a diagnostic test is used to restrict eligibility for a randomized clinical trial comparing a regimen containing a new drug to a control regimen. This approach, was used for the development of trastuzumab in which patients with metastatic breast cancer whose tumors expressed HER2 in an immunohistochemistry test were eligible for randomization. Simon and Maitournam [8–10] studied the efficiency of this approach relative to the standard approach of randomizing all patients without using the test at all. They found that the efficiency of the enrichment design depended on the prevalence of test-positive patients and on the effectiveness of the new treatment in test-negative patients. When fewer than half of the patients are test positive and the new treatment is relatively ineffective in test-negative patients, the number of randomized patients required for an enrichment design is dramatically smaller than the number of randomized patients required for a standard design. For example, if the treatment is completely ineffective in test-negative patients, then the ratio of number of patients required for randomization in the enrichment design relative to the number required for the standard design is approximately $1/\gamma^2$ where γ denotes the proportion of patients who are test positive [10]. The treatment may have some effectiveness for test-negative patients either because the assay is imperfect for measuring deregulation of the putative molecular target or because the drug has off-target antitumor effects. Even if the new treatment is half as effective in test-negative patients as in test-positive patients, however, the randomization ratio is approximately $4/(\gamma + 1)^2$. This equals about 2.56 when $\gamma = 0.25$, i.e., 25% of the patients are test positive, indicating that the enrichment design reduces the number of required patients to randomize by a factor of 2.56.

The enrichment design was used for the development of trastuzumab and led to the approval of the drug for metastatic and primary breast cancer even though the test was imperfect and has subsequently been improved. The enrichment design enabled the drug to be evaluated in the patients for whom there was a biological rationale for expecting a benefit and to avoid exposing the others to a drug with

serious toxicities. Simon and Maitournam also compared the enrichment design to the standard design with regard to the number of screened patients. Zhao and Simon have made the methods of sample size planning for the design of enrichment trials available on line at http://brb.nci.nih.gov. The web-based programs are available for binary and survival/disease-free survival endpoints. The planning takes into account the performance characteristics of the tests and specificity of the treatment effects. The programs provide comparisons to standard non-enrichment designs based on the number of randomized patients required and the number of patients needed for screening to obtain the required number of randomized patients.

The enrichment design is appropriate for contexts where there is a strong biological basis for believing that test-negative patients will not benefit from the new drug. In such cases, including test-negative patients may raise ethical concerns and may confuse the interpretation of the clinical trial. As described in the section on "stratification designs," if test-negative patients are to be included, then they must be included in sufficient numbers and the number of test-positive patients must be designed to provide adequate separate analysis of the two groups. Often this is not done and instead one sees a mixed population of patients in an inadequately sized trial leading to ambiguous conclusions.

3 Biomarker Stratified Design

When a predictive classifier has been developed but there is not compelling biological or phase II data that test-negative patients do not benefit from the new treatment, it is generally best to include both classifier positive and classifier negative in the phase III clinical trials comparing the new treatment to the control regimen. In this case it is essential that an analysis plan be predefined in the protocol for how the predictive classifier will be used in the analysis. The analysis plan will generally define the testing strategy for evaluating the new treatment in the test-positive patients, the test-negative patients and overall. The testing strategy must preserve the overall type I error of the trial and the trial must be sized to provide adequate statistical power for these tests. It is not sufficient to just stratify, i.e., balance, the randomization with regard to the classifier without specifying a complete analysis plan. The main value of "stratifying" (i.e., balancing) the randomization is that it assures that only patients with completed test results will enter the trial. Pre-stratification of the randomization is not necessary for the validity of inferences to be made about treatment effects within the test-positive or test-negative subsets. The test used in the pivotal clinical trial should be analytically validated; that is, previously demonstrated to be accurate, reproducible, and robust to sources of laboratory variation. If an analytically validated test is not available at the start of the trial but will be available by the time of analysis, then it may be preferable not to pre-stratify the randomization process but to perform the analytically validated assay later on tumor specimens collected prior to randomization.

The purpose of the pivotal trial is to evaluate the new treatment overall and in the subsets determined by the prespecified classifier. The purpose is not to modify or optimize the classifier. If the classifier is a composite gene expression-based classifier, the purpose of the design is not to reexamine the contributions of each gene. If one does any of this, then an additional phase III trial may be needed to evaluate treatment benefit in subsets determined by the new classifier. Several primary analysis plans have been described by Simon [7, 11, 12], and a web-based tool for sample size planning with these analysis plans is available at http://brb.nci.nih.gov. For example, if one does not expect the treatment to be effective in the test-negative patients unless it is effective in the test-positive patients, one might first compare treatment versus control in test-positive patients using a threshold of significance of 5%. Only if the treatment versus control comparison is significant at the 5% level in test-positive patients, the new treatment will be compared to the control among test-negative patients, again using a threshold of statistical significance of 5%. This sequential approach controls the overall type I error at 5% since a treatment ineffective in both test-negative and test-positive patients has a 5% chance of being found significant for test positives, and if that comparison is not significant, the comparison for the test negatives is not performed. To have 90% power in the test-positive patients for detecting a 50% reduction in hazard for the new treatment versus control at a two-sided 5% significance level requires about 88 events of test-positive patients, the same as for an enrichment design limited to test-positive patients. If at the time of analysis the event rates in the test-positive and test-negative strata are about equal, then when there are 88 events in the test-positive patients, there will be about $88(1-\gamma)/\gamma$ events in the test-negative patients where γ denotes the proportion of test-positive patients. If 25% of the patients are test positive, then there will be approximately 264 events in test-negative patients. This will provide approximately 90% power for detecting a 33% reduction in hazard at a two-sided significance level of 5%. In this case, the trial will not be delayed compared to the enrichment design, but a large number of test-negative patients will be randomized, treated, and followed on the study rather than excluded as for the enrichment design. This will be problematic if one does not, a-priori, expect the new treatment to be effective for test-negative patients. In this case it will be important to establish an interim monitoring plan to terminate accrual of test-negative patients when interim results and prior evidence of lack of effectiveness makes it no longer viable to enter them. Most frequentist interim monitoring plans provide insufficient protection for test-negative patients in this circumstance and Karuri and Simon have recently developed a Bayesian design based on an informative prior that reflects the a-priori degree of confidence in the test [13]. The Karuri and Simon Bayesian design protects the chance of false positive conclusions for the study overall, and for the test-positive and test-negative patients separately.

In the situation where one has limited confidence in the predictive marker it can be effectively used for a "fall-back" analysis. Simon and Wang [14] proposed an analysis plan in which the new treatment group is first compared to the control group overall. If that difference is not significant at a reduced significance level such as 0.03, then the new treatment is compared to the control group just for test-positive

patients. The latter comparison uses a threshold of significance of 0.02, or whatever portion of the traditional 0.05 not used by the initial test. If the trial is planned for having 90% power for detecting a uniform 33% reduction in overall hazard using a two-sided significance level of 0.03, then the overall analysis will take place when there are 297 events. If the test is positive in 25% of patients and the event rates in test-positive and test-negative patients are about equal at the time of analysis, then when there are 297 overall events there will be approximately 75 events among the test-positive patients. If the overall test of treatment effect is not significant, then the subset test will have power 0.75 for detecting a 50% reduction in hazard at a two-sided 0.02 significance level. By delaying the treatment evaluation in the test-positive patients power 0.80 can be achieved when there are 84 events and power 0.90 can be achieved when there are 109 events in the test-positive subset.

Wang et al. have shown that the power of this approach can be improved by taking into account the correlation between the overall significance test and the significance test comparing treatment groups in the subset of test-positive patients [15]. So if, for example, a significance threshold of 0.03 has been used for the overall test, the significance threshold for used for the subset can be somewhat greater than 0.02 and still have the overall chance of a false positive claim of any type limited to 5%. In the following descriptions of biomarker designs that use the fall-back analysis plan, we use the partition 0.03 overall analysis and 0.02 for subset analysis only for concreteness. Any partition that adds to 0.05 will preserve the type I error but sample size and power may vary substantially depending on the partition used. In many cases allocating most of the 5% to the subset analysis will be advantageous because having adequate sample size to achieve adequate power for the subset analysis is more constraining than obtaining adequate power for the overall analysis.

4 Designs That Evaluate a Small Number of Classifiers

The prospective drug and companion diagnostic test approach is being used today in the development of many new cancer drugs where the biology of the drug target is well understood. Because of the complexity of cancer biology, however, there are many cases in which the biology of the target is not well understood at the time that the phase III trials are initiated. We have been developing adaptive designs for these settings. The designs are adaptive, not with regard to sample size or randomization ratio, but rather with regard to the subset in which the new treatment is evaluated relative to the control.

For example with the adaptive threshold design [16] we assumed that a predictive biomarker score was prospectively defined in a randomized clinical trial comparing a new treatment T to a control C. The score is not used for restricting eligibility and no cut-point for the score is prospectively indicated. A fall-back analysis begins as described above by comparing T to C for all randomized patients using a significance threshold α_1, say 0.03, less than the traditional 0.05. If the treatment

effect is not significant at that level, then one finds the cut-point s* for the biomarker score which leads to the largest treatment effect in comparing T to C restricted to patients with score greater than s*. Jiang et al. [16] employed a log-likelihood ratio measure of treatment effect and let L* denote the log-likelihood ratio of treatment versus control effect when restricted to patients with biomarker level above s*. The null distribution of L* was determined by repeating the analysis after permuting the treatment and control labels a thousand or more times. If the permutation statistical significance of L* is less than $0.05\text{-}\alpha_1$ (e.g., 0.02), then treatment T is considered superior to C for the subset of the patients with biomarker level above s*. Jiang et al. provided bootstrap confidence intervals for s*. They provided an approach to sample size planning for a trial based on this fallback strategy and also upon a more powerful strategy that does not utilize a portion of the total type I error for a test of the overall null hypothesis of average treatment effect.

The analysis plan used in the adaptive threshold design is based on computing a global test based on a maximum test statistic. For the adaptive threshold design, the maximum is taken over the set of cut-points of a biomarker score. The idea of using a global maximum test statistic is much more broadly applicable, however. For example, suppose multiple candidate binary tests, $B_1, \ldots, B_K$ are available at the start of the trial. These tests may or may not be correlated with each other. Let L_k denote the log-likelihood of treatment effect for comparing T to C when restricted to patients positive for biomarker k. Let L* denote the largest of these values and let k* denote the test for which the maximum is achieved. As for the adaptive threshold design, the null distribution of L* can be determined by repeating the analysis after permuting the treatment and control labels a thousand or more times. If the permutation statistical significance of L* is less than $0.05\text{-}\alpha_1$ (e.g., 0.02), then treatment T is considered superior to C for the subset of the patients positive for biomarker test k*. The stability of the indicated set of patients who benefit from T (i.e., k*) can be evaluated by repeating the computation of k* for bootstrap samples of patients.

5 Predictive Analysis of Clinical Trials

Freidlin and Simon [17] also published an adaptive signature design for settings where a single or small number of candidate classifiers are not available at the start of the phase III clinical trial. At the time of final analysis, one starts by comparing outcomes for the treatment group T to the control group C for all randomized patients. If this overall treatment effect is not significant at a reduced level α_1, the full set P of patients in the clinical trial is partitioned into training set Tr and validation set V. A prespecified algorithmic analysis plan is applied to the training set to generate a "predictive classifier" F(x;Tr) where x denotes the vector of variables available. This vector may include only candidate classifiers or may include variables with no a-priori credentials for predictive classification.

The design was originally proposed for settings where the x vector included gene expression values from a genome-wide expression measurement. A predictive classifier is a function that identifies the patients who appear to benefit from the new treatment T compared to the control C; F(x;Tr) = 1 means that a patient with covariate vector x is predicted to benefit from T whereas F(x;Tr) = 0 indicates that patient is not predicted to benefit from T. This is a predictive classifier based on comparing two treatment groups, not the more familiar kind of prognostic classifier for a single group. This classifier is developed based on analyzing outcome and covariates for the two treatment groups in the training set. Freidlin and Simon developed a weighted voting predictive classifier based on genes whose expression levels indicate an interaction with treatment in predicting outcome. Many other types of classifier development algorithms are possible and the design can be used broadly, not just when the covariates represent gene expression measurements. For example with survival data one could use a proportional hazards model

$$\log \frac{h(t;\underline{x},z)}{h_0(t)} = \alpha z + \underline{\beta}'\underline{x} + z\underline{\eta}'\underline{x}$$

where z is a treatment indicator z = 0 for C and z = 1 for T and $\underline{x}$ denotes the vector of covariates. This model can be fit on the training set by maximizing the penalized log partial likelihood with an L1 penalty on the components of the main effect vector $\underline{\beta}$ and the treatment by interaction vector $\underline{\eta}$. The difference in log hazard for a patient with covariate vector $\underline{x}$ receiving treatment T compared to that same patient receiving treatment C is estimated by $\delta(\underline{x}) = \hat{\alpha} + \hat{\underline{\eta}}'\underline{x}$. This function can be used to classify or rank patients in the validation set. Patients with the most negative values of $\delta(\underline{x})$ are predicted to be the most likely to benefit from T relative to C. In order to classify patients in the validation set, a cut-point must be defined. This can either be a predetermined value such as zero, or a predetermined quantile of the distribution of $\delta(\underline{x})$ in the training set or used as an additional tuning parameter. All tuning parameters should be optimized by cross-validation within the training set.

An alternative classifier can be based on a generalization of the compound covariate method of Radmacher et al. [18]. The compound covariates are defined based on fitting single variable proportional hazards models

$$\log \frac{h(t;x_i,z)}{h_0(t)} = \alpha z + \beta_i{}' x_i + z\eta_i{}' x_i$$

for each variable i = 1,2,...,p where z denotes a treatment indicator z = 0 for C and z = 1 for T. One obtains estimates $\hat{\beta}_i$ and $\hat{\eta}_i$. The two compound covariates are defined as

$$v_1 = \sum \hat{\beta}_i x_i \text{ and } v_2 = \sum \hat{\eta}_i x_i$$

where the summations are over the variables for which the treatment by covariate interactions are nominally significant at level ξ in the corresponding univariate

models. ξ is used as a tuning parameter. Prediction is based on the proportional hazards model involving only treatment and the two compound covariates:

$$\log \frac{h(t; v_1, v_2, z)}{h_0(t)} = \alpha z + \beta^* v_1 + z\eta^* v_2 \quad (1)$$

For predicting treatment the difference in log hazard for a patient with compound covariate values (v_1,v_2) receiving treatment T compared to that same patient receiving treatment C is estimated as $\hat{\alpha} + \hat{\eta}^* v_2$. This function can be used to classify or rank patients in the validation set. Patients with the most negative values of $\hat{\alpha} + \hat{\eta}^* v_2$ are predicted to be the most likely to benefit from T relative to C. For evaluating prediction accuracy one would classify patients in the validation set into quantiles based on their values of $\hat{\alpha} + \hat{\eta}^* v_2$ and examine the actual treatment effect within those quantiles.

A similar approach to that described above is to use a model like [1] for prediction but where v_1 is defined as the first supervised principal component of expression levels for the variables that are prognostic at nominal univariate significance level ξ and v_2 being the first principal component of expression levels of the variables that have nominal ξ level significant interactions. The level ξ is a tuning parameter [19].

Once a single completely specified classifier is defined on the training set, it is used to classify the patients in the validation set. These patients are classified as either "sensitive" to the new treatment, i.e., predicted likely to benefit from the new treatment T relative to C or not sensitive. Let S denote the set of sensitive patients in the validation set; i.e., $S = \{j \varepsilon V | F(x_j, Tr) = 1\}$. One then compares outcomes for sensitive patients in the validation set who received T versus sensitive patients in the validation set who received C. Let L denote the log-rank statistic (if outcomes are time-to-event) for this comparison of T vs C of sensitive patients in the validation set. The null distribution of L is determined by repeating the entire analysis after permuting the treatment and control labels a thousand or more times. If the permutation statistical significance of L is less than $0.05\text{-}\alpha_1$ (e.g., 0.02), then treatment T is considered superior to C for the subset of the patients predicted to be sensitive using the classifier developed in the training set.

Freidlin et al. [20] demonstrated that the statistical power of this approach can be substantially increased by embedding the classifier development and validation process in a K fold cross-validation. This idea is very powerful and much more broadly applicable than in the context described by Freidlin et al. [17] The concept is to prospectively define an algorithm A for classifying patients as likely or not likely to have better outcome on the new treatment T compared to the control C. This algorithm constitutes the entire preplanned subset analysis. In contrast to the usual subset analysis which results in a bunch of statements about statistical significance of treatment effects within multiple subsets, this algorithm results in a single completely determined predictive classifier. The predictive classifier partitions the space of covariate vectors into a region for patients who are predicted to benefit from the new treatment T and the complementary region for patients who are not predicted to benefit from T. This algorithm might, for example, be defined

as indicated above by fitting a proportional hazards model involving treatment, main covariate effects and treatment by covariate interactions to the data and then defining the predictive classifier based on imposing a cut-point on the difference in log-likelihood for the predictive index computed if the new treatment is used minus if the control is used. With this approach the model could be fit to the high-dimensional data using penalized likelihood methods or univariate screening to find covariates with apparent interactions with treatment. Many other kinds of algorithms are possible. The algorithm A when applied to a dataset D defines a completely specified predictive classifier F(x|D, A). The classifier that will potentially be used in the future is the one obtained by applying the algorithm to the full dataset (P), i.e., F(x|P,A). But first it is necessary to evaluate the algorithm using cross-validation. It should be emphasized that the cross-validation procedure does not provide some abstract characteristic of the algorithm A, it provides an almost unbiased estimate of the predictive accuracy of the classifier F(x|P, A) obtained by applying A to the full set of data D.

The cross-validation is performed in the following way. The full set (P) of patients in the clinical trial is partitioned into K disjoint subsets $P_1, \ldots, P_K$. The *i*th training set T_i consist of the full set of patients except for the *i*th subset; i.e., $T_i = P\text{-}P_i$. Let $F(x|T_I, A)$ denote the binary classifier developed by applying the algorithm A to training set T_i. Use this classifier to classify the patients in the omitted subset P_i. Let $v_j = F(x_j|T_i, A)$ denote the predictive classification for patients j in P_i. $v_j = 1$ if the patient is predicted to be sensitive to the new treatment T relative to control C, and zero otherwise. Since the patients in P_i were not included in the training set T_i used to train $F(x|T_i, A)$, this classification is predictive, not just evaluating goodness of fit to the same data used to develop the classifier. Since each patient appears in exactly one P_i, each patient is classified exactly once and that classification is done with a classifier developed using a training set not containing that patient.

Let S denote the set of patients j for whom $v_j = 1$, i.e., who are predicted to be sensitive to the new treatment. We can evaluate the predictive value of our algorithm by comparing outcomes of the patients in S who received treatment T to the outcomes for the patients in S who received the control C. Let L(S) denote a measure of difference in outcomes for that comparison; e.g., a log-rank statistic if outcomes are time-to-event. We can generate an approximation to the null distribution of L by repeating the entire analysis for thousands of random permutations of the treatment labels. This test can be used as the primary significance test of the clinical trial to test the strong null hypothesis that the new treatment and control are equivalent for all patients on the primary endpoint of the trial. Alternatively, it can be used as a fall-back test as described in the previous sections.

Having rejected the null hypothesis described above, the application of the algorithm A to the full dataset P provides a decision tool F(x|P, A) that can be used by physicians for informing future treatment decisions for their patients. The classifier recommended for future use is the one obtained by applying the algorithm to the full dataset, i.e., F(x;P,A). The K-fold cross-validation provides a proper statistical significance test and provides important information about this full sample classifier. Freidlin et al. showed that the hazard ratio for T vs C in the cross-validated

set S is a conservative estimate of the hazard ratio for the sensitive set of the full sample classifier, i.e., for the set of future patients with covariate vectors for which $F(x|P, A) = 1$.

The effectiveness of the decision tool based on $F(x|P, A)$ depends on the algorithm used. Algorithms that over-fit the data will provide classifiers that make poor predictions. Algorithms based on Bayesian models with many parameters and non-informative priors may be as prone to over-fitting as frequentist models with many parameters. The effectiveness of an algorithm will also depend on the dataset, i.e., the unknown truth about how treatment effect varies among patient subsets. A strong advantage of the proposed approach, however, is that an almost unbiased estimate of the performance of a defined algorithm can be obtained from the dataset of a clinical trial itself. This can be compared to treating all patients or no patients based on the results of the conventional overall null hypothesis test. This is clearly preferable to performing exploratory analysis on the full dataset without any cross-validation, reporting the very misleading goodness of fit of the model to the same data used to develop the model, and cautioning that the results need testing in future clinical trials.

The approach described above can also be used in a clinical trial for which the overall treatment effect is significant. The approach permits one to identify, based on covariate profiles, the patients who do and do not benefit from the new treatment. Rather than just focusing on the patients predicted to be sensitive to the new treatment, one also compares treatment effects for the complementary subset defined by the cross-validated classifications. To illustrate this approach we have applied it to data from the gene expression profiling study conducted on pretreatment biopsy specimens from 181 patients with diffuse large-B-cell lymphoma (DLBCL) who received a standard chemotherapy combination called CHOP and 233 patients with this disease who received R-CHOP (CHOP plus the antibody Rituximab) was analyzed [19]. Unfortunately, this was not a randomized clinical trial, but the data will serve to illustrate the method of analysis.

The only clinical covariate considered for this analysis was the international prognostic index (IPI). For the purpose of this analysis the IPI was categorized into two groups—subjects with IPI scores of 0, 1, or 2 were categorized as "low" and subjects with IPI scores 3, 4, or 5 were called "high." For some of the subjects one or more of the variables that make up the IPI were missing. If for a given subject the value for the missing variable would not change the IPI group call (e.g., depending on the value of the missing variable the IPI value would be either 1 or 2), then the subject would be included as a member of that IPI group. However if the missing value could make a difference (e.g., between 2 and 3), then that subject was excluded from our analysis. Thirty-nine subjects were excluded because the IPI group could not be determined. Of the resulting 375 subjects 262 fell in IPI class "low" and 113 subjects were in IPI class "high." The end-point was overall survival (death from any cause). Prior to the start of analysis one subject who had a survival time of zero was removed, resulting in 374 subjects for this analysis.

Gene expression and clinical data were obtained from the Gene Expression Omnibus (acc. no. GSE10846). To account for the differences in microarray preprocessing between R-CHOP and CHOP samples, the expression values for each gene in the R-CHOP group was adjusted so that its median matched the median of

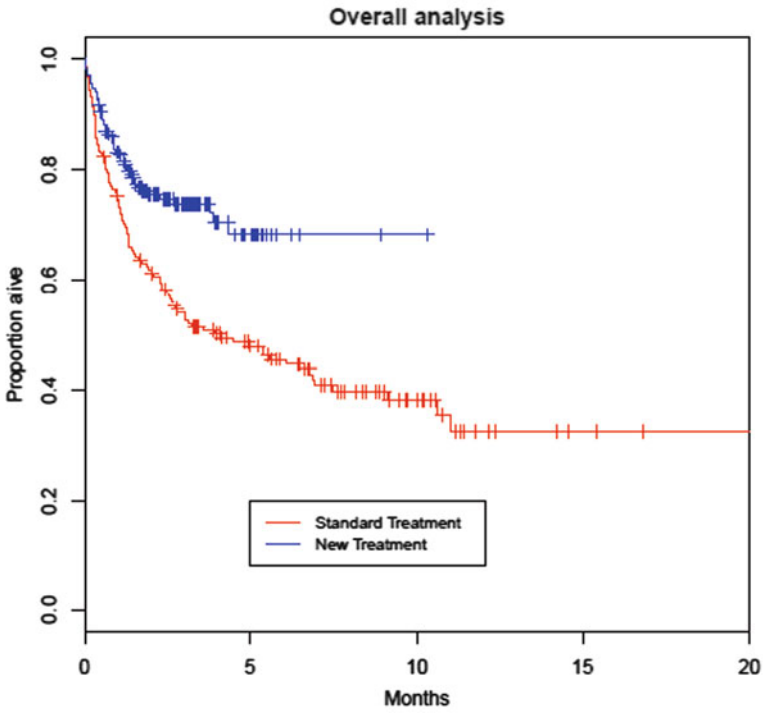

Fig. 1 Overall analysis. The value of the log-rank statistic is 14.1 and the corresponding p-value is 0.0002. The new treatment thus shows an overall benefit

the CHOP group [21]. For the predictive analysis the data were log2 transformed and the 1,000 genes with the highest variance were used.

A Cox proportional hazards (ph) regression model was developed using patients in both CHOP (standard treatment acronym of four chemotherapy drugs, C) and R-CHOP (new treatment consisting of standard CHOP plus antitumor antibody rituximab, E) groups. A 10-fold cross-validation was applied to estimate predictive accuracy. Univariate gene selection was used as the feature selection method. For each gene, a Cox ph model was developed in the training set using treatment, covariate (IPI), gene, and treatment by covariate and treatment by gene interactions. Ten genes with the lowest p-values for the treatment by gene interactions were selected for inclusion in the multivariate Cox ph model. The multivariate Cox ph model was again developed using treatment, IPI, 10 best genes, treatment by IPI and treatment by gene interactions. Gene selection and multivariate model development were all done within each cross-validation loop. In the multivariate Cox ph model, let $\delta =$ main effect of treatment, $\gamma =$ vector of interaction coefficients. Then, for a patient in the test set with covariate vector X_{test} $F(X_{test}) = 1$ if $\delta + \gamma' X_{test} < c$ where c was fixed to be the median of the F(X) values in the corresponding training set.

Figure 1 shows the results of the overall analysis. The results of applying the predictive algorithm in a ten-fold cross-validation loop are shown in Figs. 2 and 3.

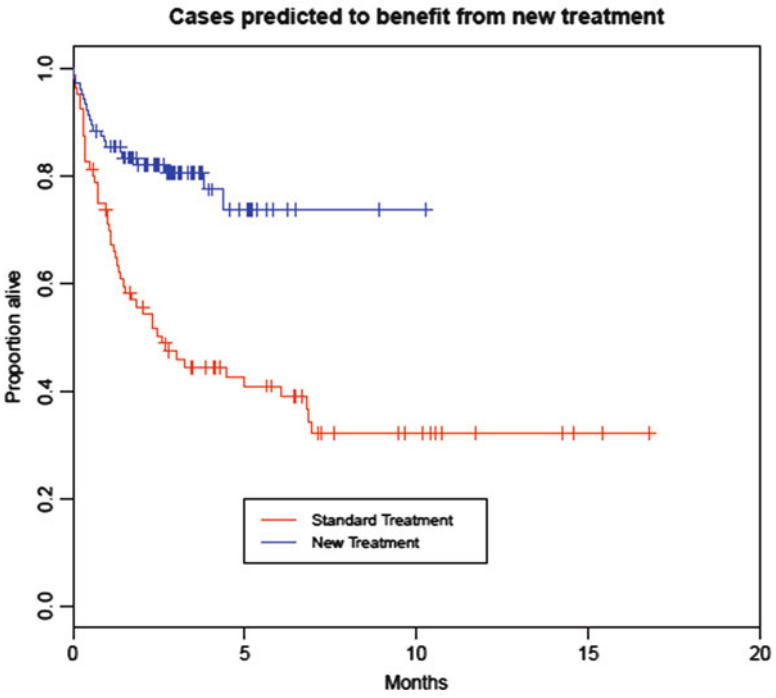

Fig. 2 Predictive analysis. Cross-validation was used to predict patients who would benefit or not from the new treatment. This figure shows the survival curves for patients predicted to benefit from the new treatment. The value of the log-rank statistic for the separation of the survival curves is 19.67 and the permutation p-value is 0.005 (200 permutations). The hazard ratio is −1.12 and the bootstrap-based 95% CI for the HR is (−1.40, −0.125) (200 bootstrap samples)

For the sensitive subset of patients who appear to benefit from R-CHOP, the value of the log-rank statistic for the separation of the cross-validated survival curves is 19.67 and the permutation p-value is 0.005 (200 permutations). For this sensitive subset, the log hazard ratio is −1.12 and the bootstrap based 95% CI for the log HR is (−1.40, −0.125) (200 bootstrap samples). For the complementary subset of patients who do not appear to benefit from R-CHOP, the value of the log-rank statistic for the cross-validated survival curves is 0.81 and the permutation p-value is 0.49 (200 permutations). The predictive analysis has thus identified a group of patients who are unlikely to benefit from the new treatment. Table 1 provides some information about the proportional hazards model developed on the full dataset. The p values listed are nominal p values conditional on including the 10 gene expression variables with the most nominally significant univariate interactions with treatment. Table 2 lists the genes that have more than 50% cross-validation support. The "%cv support" column indicates the proportion of the 10 loops of the cross-validation that the variable was selected for inclusion in the model.

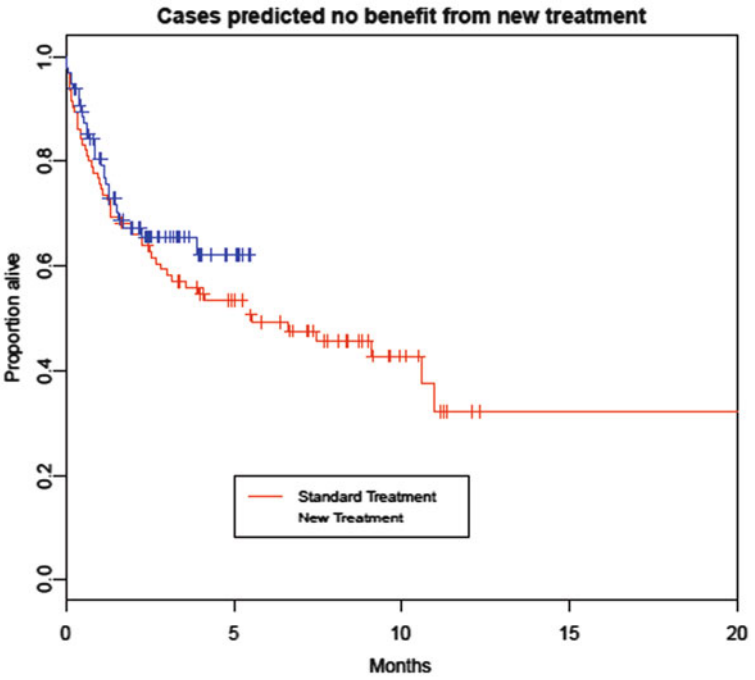

Fig. 3 Survival curves for cases predicted no benefit from new treatment. The value of the log-rank statistic in this case is 0.81 and the permutation p-value is 0.49 (200 permutations)

Table 1 The coxph model on applying the algorithm to the full dataset (for classifying future patients)

Variable	Coefficient	p-Value	Variable	Coefficient	p-Value
T	−1.69	0.27			
1552531_a_at	0.07	0.21	T*1552531_a_at	−0.15	0.06
210313_at	0.12	0.03	T*210313_at	−0.21	0.03
242334_at	−0.05	0.20	T*242334_at	−0.07	0.42
242107_x_at	−0.13	0.06	T*242107_x_at	0.35	0.002
231391_at	0.05	0.28	T*231391_at	−0.26	0.004
1565026_a_at	0.15	0.01	T*1565026_a_at	−0.34	0.0006
206413_s_at	−0.03	0.39	T*206413_s_at	0.11	0.08
203641_s_at	−0.18	0.008	T*203641_s_at	0.23	0.02
231898_x_at	−0.06	0.20	T*231898_x_at	0.20	0.007
243905_at	−0.06	0.28	T*243905_at	0.31	0.003
IPI	−1.6	<0.0001	T*IPI	0.29	0.43

T denotes treatment indicator. Variables with "at" suffix represent gene expression levels for Affymetrix probe sets. p-values are nominal values which ignore the effect of variable selection

Table 2 Genes with more than 50% cross-validation support (i.e., chosen as one of the 10 best genes in more than 5 CV loops)

Gene	% CV support	Name (symbol)	Molecular function
210313_at	100	Leukocyte immunoglobulin receptor, subfamily A, member 4 (LILRA4)	Receptor activity
1552531_a_at	100	NLR family, pyrin domain containing 11 (NLRP11)	Nucleotide binding, protein binding, ATP binding
242334_at	80	NLR family, pyrin domain containing 4 (NLRP4)	Nucleotide binding, protein binding, ATP binding
231391_at	80	Cortexin 3 (CTXN3)	Unknown
1565026_a_at	70	Orofacial cleft 1 candidate 1 (OFCC1)	Unknown
242107_x_at	70	Unknown	Unknown
206413_s_at	70	Unknown	Protein binding

6 Conclusion

Developments in genomics have increased the focus of biostatisticians on prediction problems. This has led to many useful developments for predictive modeling where the number of variables is larger than the number of cases. Heterogeneity of human diseases and new technology for characterizing diseased tissue presents new opportunities and challenges for the design and analysis of clinical trials. In oncology, treatment of broad populations with regimens that do not benefit most patients is less economically sustainable with expensive molecularly targeted therapeutics. The established molecular heterogeneity of human diseases requires the development of new paradigms for the use of randomized clinical trials as a reliable basis predictive medicine [1, 2]. We have presented here prospective designs for the development of new therapeutics with candidate predictive biomarkers. An approach to the Predictive Analysis of Clinical Trials (PACT) has also been presented. This approach preserves the type I error of the study and uses re-sampling to develop and validate a predictive classifier that can be used to inform treatment selection for future patients. This approach provides a statistically sound framework for bridging the gap between clinical trials and clinical practice that has long existed and may serve as a basis for clinical trials in the era of predictive medicine.

References

1. Simon R (2004) An agenda for clinical trials: clinical trials in the genomic era. Clin Trials 1:468–470
2. Simon R (2007) New challenges for 21st century clinical trials. Clin Trials 4:167–169
3. Peto R, Pike MC, Armitage P (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 34:585

4. Peto R, Pike MC, Armitage P (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 35:1
5. Dixon DO, Simon R (1991) Bayesian subset analysis. Biometrics 47:871
6. Sawyers CL (2008) The cancer biomarker problem. Nature 452:548–552
7. Simon R (2005) A roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23:7332–7341
8. Simon R, Maitournam A (2005) Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10:6759–6763
9. Simon R, Maitournam A (2006) Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and correction. Clin Cancer Res 12:3229
10. Maitournam A, Simon R (2005) On the efficiency of targeted clinical trials. Stat Med 24: 329–339
11. Simon R (2008) Using genomics in clinical trial design. Clin Cancer Res 14:5984–5993
12. Simon R (2010) Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Per Med 7(1):33–47
13. Karuri S, Simon R (2012) A two-stage Bayesian design for co-development of new drugs and companion diagnostics. Stat Med 31:901–914
14. Simon R, Wang SJ (2006) Use of genomic signatures in therapeutics development. Pharmacogenomics J 6:166–173
15. Wang SJ, O'Neill RT, Hung HMJ (2007) Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat 6:227–244
16. Jiang W, Freidlin B, Simon R (2007) Biomarker adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst 99: 1036–1043
17. Freidlin B, Simon R (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 11:7872–7878
18. Radmacher MD, McShane LM, Simon R (2002) A paradigm for class prediction using gene expression profiles. J Comput Biol 9:531–537
19. Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. J Am Stat Assoc 101:119–137
20. Freidlin B, Jiang W, Simon R (2010) The cross-validated adaptive signature design for predictive analysis of clinical trials. Clin Cancer Res 16(2):691–698
21. Lenz G, Wright GW, Dave SS, Xiao W, Powell J, Zhao H et al (2008) Stromal gene signatures in large-B-cell lymphomas. N Engl J Med 359(22):2313–2323