

Clinical trials for predictive medicine

Richard Simon^{*†}

Developments in biotechnology and 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 the new regimen. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: clinical trials; predictive medicine; predictive biomarkers; re-sampling methods

1. Introduction

Developments in biotechnology and genomics have provided tools for understanding the biological basis for heterogeneity of tumors of the same primary site and identification of important molecular targets. It has become clear that cancers of the same primary site and stage are diverse in terms of their pathogenesis, natural history, and responsiveness to therapy; they are in many cases different diseases. Large clinical trials to identify small average treatment effects in heterogeneous groups of patients have resulted in practice standards in which many patients are treated with toxic and expensive drugs to which they do not benefit. Biostatisticians now have the opportunity to develop new approaches to clinical trial design and analysis that enables a new era of predictive medicine in which appropriate treatments can be matched to appropriate patients in a reliable manner.

The new understanding of the heterogeneous nature of tumors of the same primary site leads to new challenges for clinical trial design. The standard paradigm for the design of Phase III oncology clinical trials involves broad eligibility criteria and basing conclusions on the test of the overall null hypothesis that the average treatment effect is zero. The emphasis on broad eligibility criteria has been based on 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 the overall null hypothesis was based on concerns about the multiple testing involved in the commonly practiced exploratory *post hoc* subset analysis and the assumption that qualitative interactions are unlikely [3, 4]. The advice was to perform subset analyses, but do not believe them, and the famous subset analysis of

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC7434, Bethesda MD 20892–7434, U.S.A

*Correspondence to: Richard Simon, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC7434, Bethesda MD 20892–7434, U.S.A.

†E-mail: rsimon@nih.gov

the ISIS-2 trial based on patient astrological sign is still prominent in the minds of many statisticians [5]. This paradigm was based on implicit assumptions that qualitative interactions are unlikely 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.

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 [6], the value of clinical trials for treating individual patients has sometimes been questioned. Today we have powerful tools for characterizing the tumors biologically and can use this characterization to provide a stronger 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 some 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 [7]. It is clear that the primary analysis of the new generation of oncology clinical trials must consist of more than just testing the null hypothesis of no average effect. However, 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 [8]. 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. However, for many new oncology drugs it is the 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 to risks of adverse events but provide 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 [8]. This approach, which is being used extensively today in oncology involves: (i) 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. (ii) 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. (iii) 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. In the above description, 'completely specified' does not mean that the test is perfect. It just means that all aspects of it, including the analytes to measure and the cut-points to be used are defined. 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 biomarkers. We will introduce a paradigm, the predictive analysis of clinical trials (PACT), which permits classifier development and evaluation to be validly performed in the same clinical trial. By moving from 'subset analysis' to 'classifier development' the problem is moved from one of multiple testing to one of evaluation of classification and prediction. By careful use of cross-validation, the principle of separating classifier development from classifier evaluation can be preserved with both objectives performed within the same clinical trial. With the PACT approach, the analysis plan is carefully prespecified to ensure that treatment effects in classifier based subsets are unbiasedly estimated and that study-wise Type I error is preserved.

2. 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, shown in Figure 1, 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 [9–11] 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 standard design relative to the number required for the enrichment design is approximately $1/\gamma^2$ where γ denotes the proportion of patients who are test positive [11]. This equals a factor of 4 when half the patients are test positive. Under these conditions the enrichment design also results in a more appropriate conclusion than the standard design. The conclusion for the enrichment design would be that the treatment is effective for test positive patients and the labeling indication would be restricted to test positive patients. A standard design that was large enough to reject the null hypothesis would result in a broad approval of the treatment and would result in the ineffective treatment of test negative patients with the new drug.

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, the randomization ratio is approximately $4/(\gamma + 1)^2$. This equals about 2.56 when $\gamma = 0.25$, that is, 25% of the patients are test positive, indicating that the enrichment design reduces the number of randomized patients by a factor of 2.56.

The enrichment design was very effective for the development of trastuzumab. Even though the test was imperfect and has subsequently been improved, it was well defined and sufficiently accurate to dramatically improve the efficiency of the Phase III clinical trial relative to a design that did not utilize the test at all. Zhao and Simon have made the methods of sample size planning for the design of enrichment trials available online 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 nonenrichment 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.

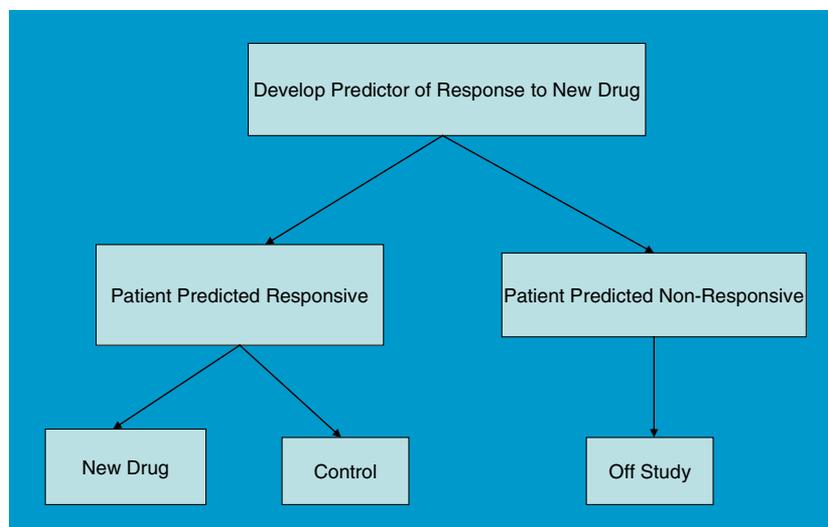


Figure 1. Enrichment design.

The enrichment design is appropriate for contexts where there is such a strong biological basis for believing that test negative patients will not benefit from the new drug that including them in would raise ethical concerns. In many situations, the biological basis is strong but not compelling. The enrichment design does not provide data on the effectiveness of the new treatment compared with control for test negative patients. Consequently, unless there is compelling biological or Phase II data that the new drug is not effective in test negative patients, the enrichment design may not provide adequate data for regulatory approval of the test.

3. Biomarker stratified design

When a predictive classifier has been developed but there is no 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 as shown in Figure 2. Although this is often referred to as the ‘stratification design’, what is really essential is that an analysis plan be predefined in the protocol for how the predictive classifier will be used in the analysis and that the trial be sized for this analysis plan. It is not sufficient to just stratify, that is, balance, the randomization with regard to the classifier without specifying a complete analysis plan or sizing the trial appropriately. The main value of ‘stratifying’ (i.e. balancing) the randomization is that it assures that to be randomized a patient must have an evaluable test performed on his/her tumor. Prestratification 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. 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 prestratify the randomization process.

With the stratification design the purpose of the clinical 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. Several primary analysis plans are presented below to illustrate that the plan should stipulate in detail how the predictive biomarker will be used in the analysis and that there should be no exploratory aspect to the treatment evaluation. These strategies are discussed in greater detail by Simon [12, 13] and a web-based tool for sample size planning with these analysis plans is available at <http://brb.nci.nih.gov>.

3.1. Analysis plan for biomarker with strong credentials

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 will the new treatment be compared with the control among test negative

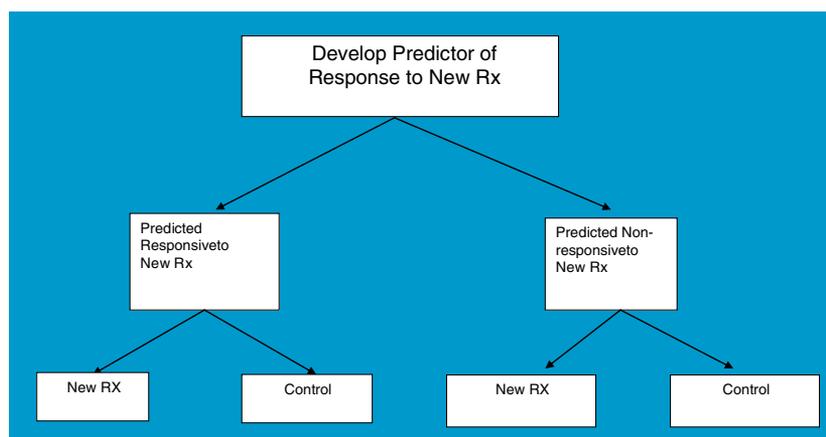


Figure 2. Biomarker stratified design.

patients, again using a threshold of statistical significance of 5%. This sequential approach controls the overall Type I error at 5%.

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. 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 with 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.

3.2. Fallback analysis plan

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 with 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 with 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 time of final analysis of 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]. Therefore, 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%.

3.3. Interaction analysis plan

A third possible analysis plan is to decide whether to compare the treatments overall or within the test positive and test negative subsets based on a preliminary test of interaction. The interaction test comparing the treatment effects for those two subsets should be one-sided, and performed at a threshold above the traditional 5% level. The larger sample size usually needed for testing for interaction at a two-sided 5% level cannot generally be justified as it requires exposing an excessive number of test negative patients to a treatment from which they are unlikely to benefit. In the example described above, the interaction test will have approximately 93.7% power at a one-sided significance level of 0.10 for detecting an interaction with 50% reduction in hazard for test positive patients and no treatment effect in test negative patients. Detailed results for this analysis plan are available using the web-based program described earlier.

4. Designs that explore 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 Jiang *et al.* [16] described the ‘adaptive threshold design’ for settings where a single predictive biomarker candidate is available but no threshold of positivity for the marker is predefined. A randomized clinical trial comparing a new treatment E to a control C is performed. The score is measured in all patients but is not used for restricting eligibility. I will present here a modification of the global test of the null hypothesis used by Jiang *et al.* Assume that there are K candidate threshold values b_1, \dots, b_K and let l_k denote the log likelihood ratio of treatment effect for the patients with biomarker value $\geq b_k$. Let b_{k^*} denote the threshold for which these log likelihood ratios are maximized and let l_{k^*} denote the maximum value. The null distribution of l_{k^*} was approximated by repeating the analysis after permuting the treatment and control labels several thousand times. If the permutation statistical significance of l_{k^*} is less than 0.05 then the global null hypothesis that treatment E is completely ineffective is rejected. Jiang *et al.* provided bootstrap confidence intervals for the threshold above which treatment E is effective. They also provided approaches to sample size planning.

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. This approach can also be used when one has a set of candidate binary biomarkers and one wishes to test whether there is a treatment effect in a subset determined by a positive value of any single marker.

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 E 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 a training set Tr and a validation set V. A prespecified algorithmic analysis plan is applied to the training set to generate a classifier $Cl(x;Tr)$. This classifier is a function that identifies the patients who appear to benefit from the new treatment. $Cl(x;Tr)=1$ means that a patient with covariate vector x is predicted to benefit from E whereas $Cl(x;Tr)=0$ indicates that a patient is not predicted to benefit from E. 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 E and \underline{x} denotes the vector of covariates. This model can be fit by maximizing the penalized log partial likelihood with separate penalties on the components of the main effect vector $\underline{\beta}$ and the treatment by interaction vector $\underline{\eta}$. The difference in hazard for a patient with covariate vector \underline{x} receiving treatment E compared with 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 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 E relative to C.

Once a single completely specified classifier is defined on the training set, it is used to classify the patients in the validation set as either ‘sensitive’ or ‘insensitive’ to treatment E. Let \mathbf{S} denote the set of sensitive patients in the validation set; that is, $\mathbf{S} = \{j \in V | Cl(\mathbf{x}_j; Tr) = 1\}$. One then compares outcomes for these sensitive patients in the validation set who received E to the sensitive patients in the validation set who received C. If the statistical significance value for this comparison is less than $0.05 - \alpha_1$ (e.g., 0.02), then treatment E 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.* [18] recently demonstrated that the 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.* [18] Although the details of this approach will not be repeated here, the concept is to prospectively define an algorithm for developing a predictive classifier using covariate vectors, treatment indicators, and outcomes for patients represented in a set of data. Applying the algorithm to the full set of patients treated in the clinical trial provides a classifier that can be used for future patients. We call this the 'indication classifier'. A conservative estimate of the treatment effect for future classifier positive patients is obtained by employing a K -fold cross-validation procedure. The classifier development algorithm is applied to K cross-validated training sets and used to classify the patients in the corresponding K cross-validated validation sets. When that is completed, all patients in the trial have been classified. The two treatment groups are then compared in the subset of classifier positive patients and the treatment effect estimated. Using that estimated treatment effect as a test statistic, its null distribution is approximated by permuting treatment labels and repeating the entire procedure thousands of times. Although computationally intensive, it is feasible on a desktop computer.

If the cross-validated treatment effect in classifier positive patients is statistically significant, the indication classifier recommended for future use is the one obtained by applying the algorithm to the full dataset. The K -fold cross-validation provides a proper statistical significance test and provides an estimate of treatment effect for this full sample classifier. Freidlin *et al.* showed that the cross-validated hazard ratio in the classifier positive subset is a conservative estimate of the hazard ratio for the sensitive set of the full sample classifier.

Applying the prespecified algorithm to bootstrap samples of the patients P in the trial provides information about the stability of the subset who benefit from the new treatment. Although the precision of the identification of this sensitive subset will be limited by the size of the clinical trial, information about specificity of treatment benefit may be substantially greater than with standard methods in which all future patients are presumed to benefit or not benefit from one treatment or the other.

The effectiveness of the indication classifier depends on the algorithm used. Algorithms that over-fit the data will provide classifiers that make poor predictions and this will be reflected in the cross-validated estimate of treatment effect for classifier positive patients. Algorithms based on Bayesian models with many parameters and noninformative priors may be as prone to over-fitting as frequentist models with too many parameters. The effectiveness of an algorithm will also depend on the dataset, that is, 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 treatment effect in future classifier positive patients can be obtained from the dataset of a clinical trial itself. 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.

This approach to PACT will be illustrated using data from an old clinical trial in which 506 patients with prostate cancer were randomly allocated to be treated with either placebo, 0.2 mg of diethylstilbestrol (DES), 1.0 mg DES, or 5.0 mg DES (19). The two lower doses (placebo and 0.2 mg DES) and the two higher doses (1.0 mg DES and 5.0 mg DES) were combined for analysis. The end-point was overall survival (death from any cause). Covariates considered for this analysis were (i) age: in years; (ii) performance status (pf) normal activity (1) versus less than normal (0); (iii) size of primary tumor in (cm^2) (sz); (iv) stage-histologic grade Index (sg); and (v) serum phosphatic acid phosphatase level (ap)

After removing records with missing observations in any of these covariates, 485 observations remained. The classifier development algorithm involved simply fitting a proportional hazards regression model including main effects of treatment and all five covariates and treatment by covariate interactions for each covariate. For any training set, let $\hat{\alpha}$ and $\hat{\eta}$ denote the maximal partial likelihood estimates of main effect of treatment and vector of five treatment by covariate interaction coefficients. For any patient with covariate vector \underline{x} , the difference between the estimated log hazard ratio if the patient receives treatment E and the estimated hazard ratio if the patient receives treatment C is $\hat{\alpha} + \hat{\eta}'\underline{x}$. The estimated main effects drop out of this difference in predictive index values. For each loop of the cross-validation, a patient in the validation set was classified as sensitive to E if $\hat{\alpha} + \hat{\eta}'\underline{x} \leq c$ where c was the median of these values for the training set.

Figure 3 shows survival curves for the overall set of patients. These curves represent the conventional analysis and the logrank significance value is 0.09. The results of applying the predictive algorithm

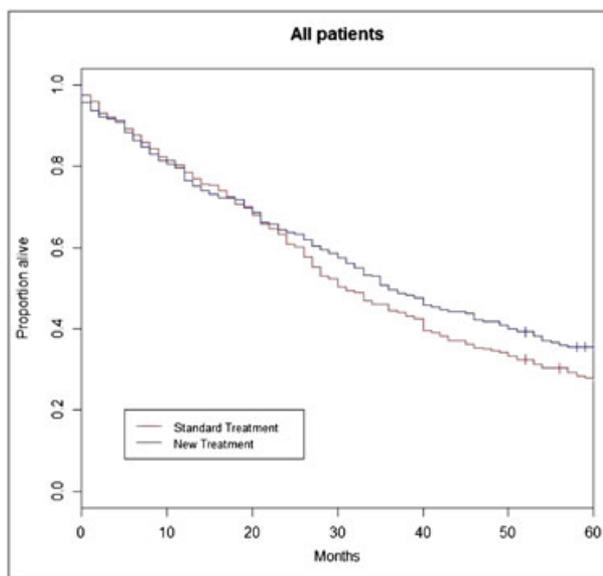


Figure 3. Comparison of survivals by treatment for prostate cancer patients over-all [19].

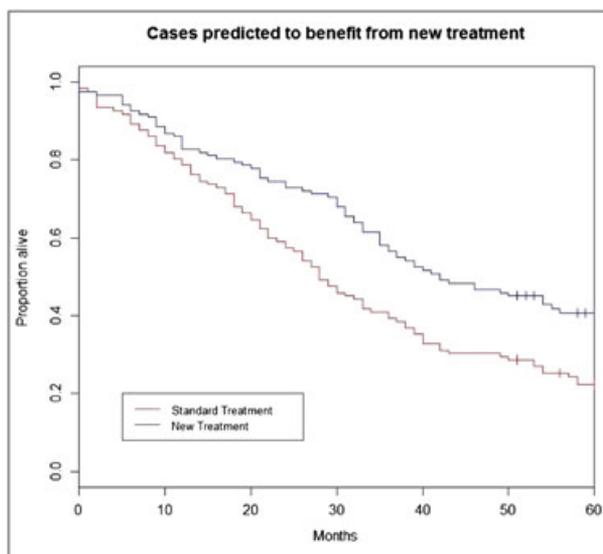


Figure 4. Comparison of survivals by treatment for prostate cancer patients classified as likely to benefit from hormonal treatment based on cross-validated classifiers.

in a 10-fold cross-validation loop are shown in Figures 4 and 5. Figure 4 shows the Kaplan–Meier survival curves by treatment for patients classified as likely to benefit from E in the 10 loops of the cross-validation. Figure 5 shows the curves for the complementary subset predicted as less likely to benefit from the high dose regimens. On the basis of 500 (more are desirable) permutations of the treatment labels (and repeating the complete cross-validation procedure for each permutation), the survival curves in Figure 4 are statistically significantly different ($p = 0.002$). Survival for patients in Figure 5 receiving the higher dose hormonal treatments are worse than for the control group although the difference is not statistically significant by permutation analysis.

Applying the model development algorithm to the full dataset gives a proportional hazards model that can be used for informing treatment selection for future patients. The estimates of regression coefficients for that model are shown below. The nominal statistical significance value for each of those coefficients is shown although those p values play no role in the use of this model for predictive analysis.

Effect	Estimated coefficient	Nominal p -value
Treatment	-2.195	0.12
Age	0.002	0.85
Performance status (pf)	-0.260	0.25
Size (sz)	0.020	0.001
Stage-grade (sg)	0.113	0.004
Acid phosphatase (ap)	0.002	0.21
Treatment*age	0.050	0.003
Treatment*pf	-0.743	0.026
Treatment*sz	-0.010	0.26
Treatment*sg	-0.074	0.19
Treatment*ap	-0.003	0.11

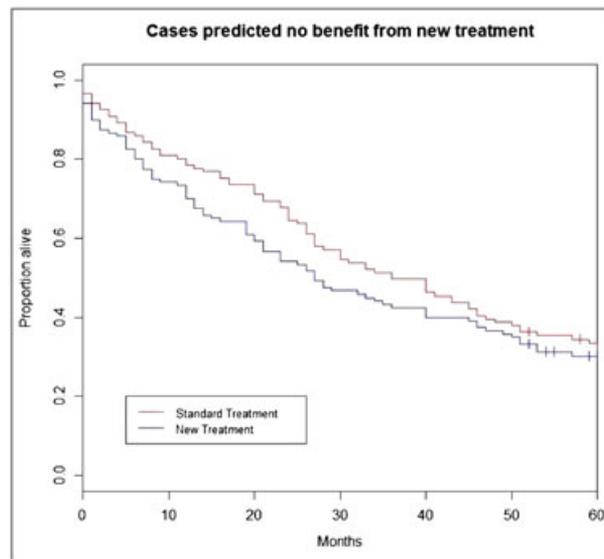


Figure 5. Comparison of survivals for treatment for prostate cancer patients classified as unlikely to benefit from hormonal treatment based on cross-validated classifiers.

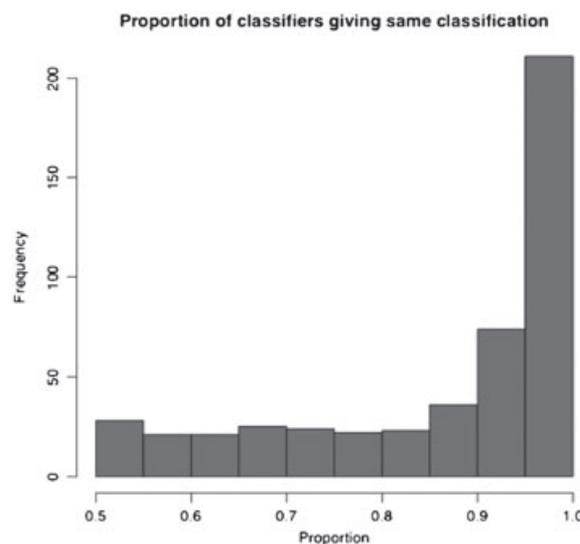


Figure 6. A predictive classifier was developed for each bootstrap sample of cases of prostate cancer data. The proportion of indication classifiers giving the same classification (likely to benefit from experimental treatment or not likely to benefit) was determined for each patient. The proportion is an indication of the stability of the classification. Figure shows a histogram of these proportions.

The hazard ratio for high dose versus control treatment increases with age and decreases with improved performance status ($pf = 1$ denotes normal activity). That is, the higher dose regimens tend to benefit younger patients and patients with good performance status. The median value of $\hat{\alpha} + \hat{\eta}'x$ for the full dataset is -0.134 . The full dataset classifier was redeveloped based on bootstrap samples to estimate the stability of classification. Figure 6 shows that classification is very stable for the majority of patients. These observations are consistent with the interpretation previously published by Byar *et al.* that higher doses of diethylstilbestrol are contraindicated for older patients because of cardiovascular toxicity [19].

6. Conclusion

Developments in biotechnology and 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]. Prospective designs for the development of new therapeutics with candidate predictive biomarkers have been presented here including an approach to the PACT. This approach preserves the Type I error of the study and uses resampling 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.

Acknowledgement

I wish to acknowledge Dr. Jyothi Subramanian for the analysis of the prostate cancer example.

References

1. Simon R. An agenda for clinical trials: Clinical trials in the genomic era. *Clinical Trials* 2004; **1**:468–470.
2. Simon R. New challenges for 21st century clinical trials. *Clinical Trials* 2007; **4**:167–169.
3. Peto R, Pike MC, Armitage P. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. *British Journal of Cancer* 1976; **34**:585–612.
4. Peto R, Pike MC, Armitage P. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II Analysis and examples. *British Journal of Cancer* 1977; **35**:1–39.
5. ISIS-2 Collaborative Group. Randomised trial of IV streptokinase, oral aspirin, both or neither among 17187 cases of suspected acute myocardial infarction. *Lancet* 1988; **2**:349–360.
6. Dixon DO, Simon R. Bayesian subset analysis. *Biometrics* 1991; **47**:871–881.
7. Sawyers CL. The cancer biomarker problem. *Nature* 2008; **452**:548–552.
8. Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* 2005; **23**:7332–7341.
9. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 2005; **10**:6759–6763.
10. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: Supplement and Correction. *Clinical Cancer Research* 2006; **12**:3229.
11. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 2005; **24**:329–339.
12. Simon R. Using genomics in clinical trial design. *Clinical Cancer Research* 2008; **14**:5984–5993.
13. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. *Expert Review of Molecular Diagnostics* 2008; **2**(6):721–729.
14. Simon R, Wang SJ. Use of genomic signatures in therapeutics development. *The Pharmacogenomics Journal* 2006; **6**:1667–1673.
15. Wang SJ, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 2007; **6**:227–244.
16. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute* 2007; **99**:1036–1043.
17. Freidlin B, Simon R. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 2005; **11**:7872–7878.
18. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design for predictive analysis of clinical trials. *Clinical Cancer Research* 2010; **16**(2):691–698.
19. Byar D. Treatment of prostatic cancer: Studies by the Veterans Administration Cooperative Urological Research Group. *New York Academy of Medicine* 1972; **48**(5):751–766.