Conducting Definitive Phase III Trials Using Pharmacogenomcs

Richard Simon
Chief, Biometric Research Branch
National Cancer Institute
http://linus.nci.nih.gov/brb

“If new refrigerators hurt 7% of customers and failed to work for another one-third of them, customers would expect refunds.”

BJ Evans, DA Flockhart, EM Meslin Nature Med 10:1289, 2004

- Clinical trial for patients with breast cancer, without nodal or distant metastases, Estrogen receptor positive tumor
  - 5 year survival rate for control group (surgery + radiation + Tamoxifen) expected to be 90%
  - Size trial to detect 92% survival in group treated with control modalities plus chemotherapy

Conditions Frequently Preceding a Major Restructuring in an Industry

- Economic stresses which cast doubt on business as usual
- Scientific progress providing new ways of doing business

- Better targeted therapies offer improved health quality and reduced waste of resources
- We need regulatory policies that encourage better targeting of therapies

Using Genomics in Development of a New Therapeutic

- Develop a pharmacogenomic classifier

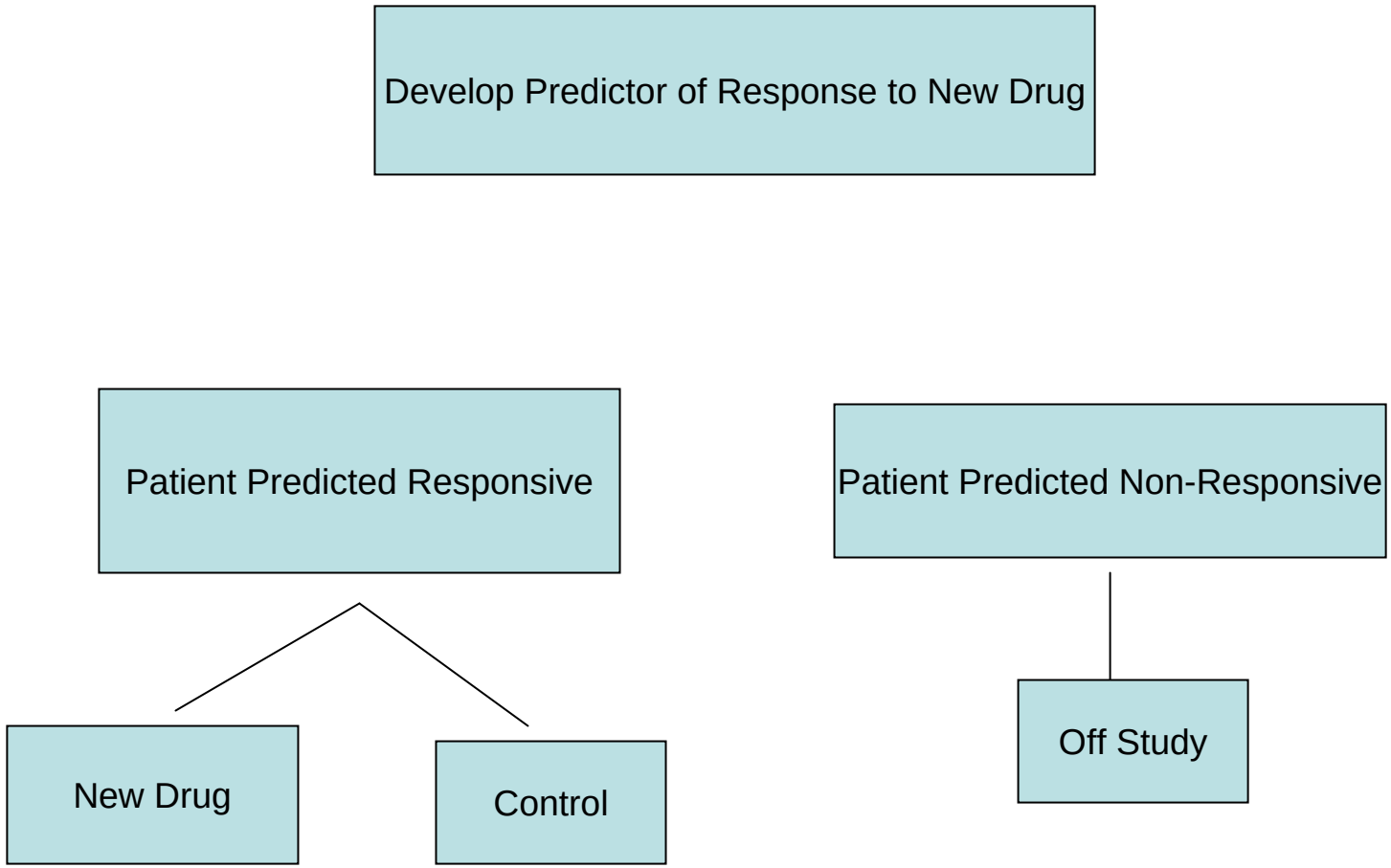
- Use a completely specified classifier developed on one set of data to obtain definitive results about effectiveness of a new treatment in a well defined population of patients
  - Use of genomics in a hypothesis testing framework
  - Avoid endless exploratory analyses that never result in reliable results

Pharmacogenomic Classifier
Composite Biomarker
Genomic Signature

- A set of genes is not a classifier

Using Genomics in Development of a New Therapeutic (I)

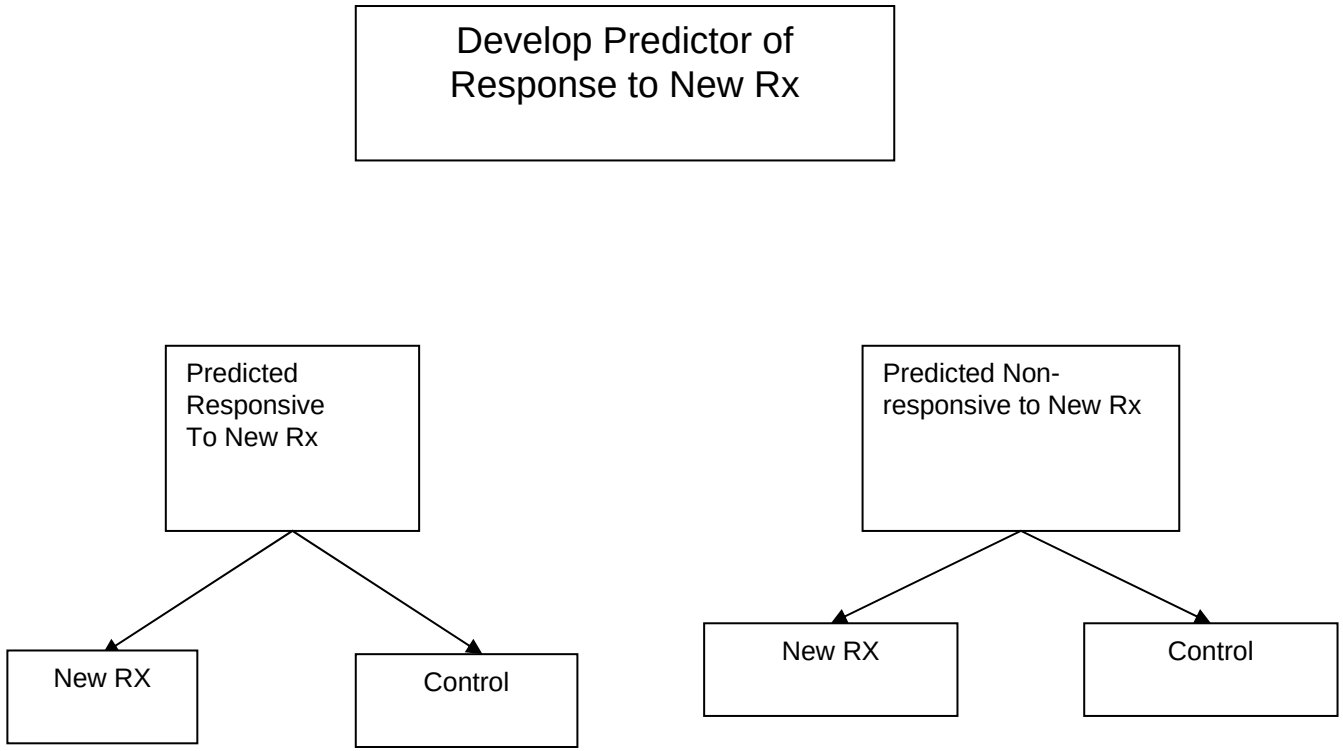
- Develop a diagnostic classifier that identifies the patients likely to benefit from the new drug
- Use the diagnostic as eligibility criteria in a prospectively planned evaluation of the new drug
- Demonstrate that the new drug is effective in a prospectively defined set of patients determined by the diagnostic
- Demonstrate that the diagnostic can be reproducibly measured
- Confirmatory phase III trial


Develop Predictor of Response to New Drug
Patient Predicted Responsive
Patient Predicted Non-Responsive
New Drug
Control
Off Study

Using Genomics in Development of a New Therapeutic (II)

- Develop a diagnostic classifier that identifies the patients likely to benefit from the new drug
- Perform separate adequate randomized clinical trials for classifier + and classifier – patients.
- Demonstrate that the diagnostic can be reproducibly measured
- Confirmatory phase III trial

Using PG Classifiers to Select Patients for Phase III Trials



Using Genomics in Development of a New Therapeutic (III)

- Develop a diagnostic classifier that identifies the patients likely to benefit from the new drug
- Do not use the diagnostic to restrict eligibility, but rather to structure a prospectively planned analysis strategy of a randomized trial of the new drug.
- Compare the new drug to the control overall for all patients ignoring the classifier.
  - If the treatment effect on the primary pre-specified endpoint is significant at the 0.04 level, then claim effectiveness for the eligible population as a whole.
- If the overall test is not significant at the 0.04 level, then perform a single subset analysis evaluating the new drug in the classifier + patients.
  - If the treatment effect is significant at the 0.01 level, then claim effectiveness for the classifier + patients.
- Demonstrate that the diagnostic can be reproducibly measured
- Confirmatory phase III trial

These Strategies Require

- The data used to develop the classifier must be distinct from the data used to test hypotheses about treatment effect in subsets determined by the classifier
  - Developmental studies are exploratory
  - Studies on which treatment effectiveness claims are to be based should be hypothesis testing studies based on completely pre-specified classifiers

Developmental Studies

- May be based on data from phase II trials or “failed” phase III trials
- May be staged to refine classifiers before use in phase III evaluations
- The objective is to develop a classifier that can be reliably measured and used to focus phase III evaluations

Phase III Treatment Evaluations in Classifier Determined Subsets

- Phase III trials of new patients in which PG classifier is measured prior to randomization

- Previously conducted randomized phase III trials in which specimens were archived

Phase III Treatment Evaluations in Previously Conducted Randomized Phase III Trials

- Data not used in development of classifier
- Prospective analysis plan based on completely specified classifier
- Completeness of specimen archive
  - What percentage of patients would not agree to specimen collection in new trial?

Genomic Classifiers Used for Targeting Patients in Drug Development

- The classifier can be considered a composite biomarker, but the components should not have to be “valid disease biomarkers” in the FDA sense

- “… a pharmacogenomic test result may be considered a valid biomarker if (1)it is measured in an analytical test system with well-established performance characteristics and (2)there is an established scientific framework or body of evidence that elucidates the physiologic, pharmacologic, toxicologic, or clinical significance of the test results.”

- “…distinction between known valid biomarkers that have been accepted in the broad scientific community and probable valid biomarkers that appear to have predictive value for clinical outcomes, but may not yet be widely accepted or independently verified by other investigators or institutions.”

Biomarker

- “Any biological measurement that provides actionable information regarding disease progression, pharmacology, or safety that can be used as a basis for decision making in drug development.”
  - J. Boguslavsky

- “I don’t know what ‘clinical validation’ [of a biomarker] means. The first thing you have to do is define a purpose for the biomarker. Validation is all about demonstrating fitness for purpose.”
  - Dr. Stephen Williams, Pfizer

Developing Composite Genomic Classifiers

- Classifiers should classify accurately
- Composite classifiers incorporate the contributions of multiple single-gene features
- The single gene feature are usually selected based on their “informativeness” for distinguishing patients likely to respond to the new rx from patients not likely to respond
- The single gene features can be selected based on informativeness in identifying patients more likely to respond to a new treatment than to a control treatment

Developing Composite Genomic Classifiers

- Classifiers should classify accurately
- To classify accurately, it is much more important that informative features not be excluded
- To classify accurately, it is less important that noise features be excluded
- If we wished to “validate” a classifier, we should validate it’s predictions, not that the same features (genes) are included in a classifier developed on independent data

After Developing the Classifier Comes

- Validation of the classifier?
- Use of the classifier to focus evaluation of the new treatment?

Development of PG Classifier of Tumor Sensitivity to Drug

- Can immensely improve the efficiency of phase III trials
  - Select patients predicted to be most sensitive
- Enables patients to be treated with drugs that actually work for them
- Avoids false negative trials for heterogeneous populations
- Avoids erroneous generalizations of conclusions from positive trials

Tumors of a given primary site are often heterogeneous with regard to oncogenesis. These tumors may represent different diseases

- “Hypertension is not one single entity, neither is schizophrenia. It is likely that we will find 10 if we are lucky, or 50, if we are not very lucky, different disorders masquerading under the umbrella of hypertension. I don’t see how once we have that knowledge, we are not going to use it to genotype individuals and try to tailor therapies, because if they are that different, then they’re likely fundamentally … different problems…”
  - George Poste

Strategies for Development of a Genomic Classifier

- During phase I/II development
  - Extended phase II
- After failed phase III trial using archived specimens
- “Prospectively” during phase III

Strategies for Identifying the Tumors for Which a Drug is Active

- Compare responders to non-responders with regard to
  - Expression of target protein
    - Herceptin
  - Mutations in target gene
    - Iressa
  - Genome-wide expression profile
  - Germline polymorphisms in candidate metabolic genes

- For Herceptin, even a relatively poor assay enabled conduct of a targeted phase III trial which was crucial for establishing effectiveness
- In many cases, the assay based on the presumed mechanism of action will not correlate with response and it may be more effective to let the data develop the assay via expression profiling

Developing a Composite Biomarker Classifier

- Feature (gene) selection
  - Which genes will be included in the model
- Select model type
- Training the model (e.g. fitting the parameters)

Most Statistical Methods Are For Inference, Not Prediction and Particularly Not for p>>n Prediction Problems

- Development and validation of diagnostic classifiers are primarily problems of prediction, not of inference about parameters
  - Predictive accuracy, not false positive genes
- Demonstrating goodness of fit of a model to the data used to develop it is not a demonstration of predictive accuracy
  - With p>>n, perfect goodness of fit is always possible
- Many standard statistical methods are not relevant or effective for p>>n prediction problems

Feature Selection Using DNA Microarray Expression Profiles

- Genes that are univariately differentially expressed among the classes at a significance level α (e.g. 0.01)
  - The α level is selected to control the number of genes in the model, not to control the false discovery rate
  - The accuracy of the significance test used for feature selection is not of major importance because identifying differentially expressed genes is not the ultimate objective

Linear Classifiers for Two Classes

$$l(\underline{x}) = \sum_{i \varepsilon F} w_i x_i$$

$\underline{x}$ = vector of log ratios or log signals

F = features (genes) included in model

w_i = weight for i'th feature

decision boundary $l(\underline{x})$ > or < d

Linear Classifiers for Two Classes

- Fisher linear discriminant analysis
- Diagonal linear discriminant analysis
- Compound covariate predictor
- Golub's weighted voting method
- Perceptrons
- Naïve Bayes classifier
- Partial least squares classifier
- Principal components classification
- Supervised principal components
- Support vector machine with inner product kernel

- When p>>>n, a linear classifier can almost always be found which fits the data perfectly.
- Why consider more complex models?
- The full set of linear models is too rich
- Restricted linear classifiers which do not attempt to minimize training error perform better by:
  - Incorporating influence of multiple variables without attempting to select the best small subset of variables
  - Do not attempt to model the multivariate interactions among the predictors and outcome

Myth

- That complex classification algorithms such as neural networks perform better than simpler methods for class prediction.

- Artificial intelligence sells to journal reviewers and peers who cannot distinguish hype from substance when it comes to microarray data analysis.
- Comparative studies have shown that simpler methods work as well or better for microarray problems because the number of candidate predictors exceeds the number of samples by orders of magnitude.

Perspective

Evaluating the Efficiency of Targeted Designs for Randomized Clinical Trials

Richard Simon and Aboubakar Maitournam
Biometric Research Branch, National Cancer Institute, Bethesda, Maryland

ABSTRACT

Purpose: **Genomic technologies make it increasingly possible to identify patients most likely to benefit from a molecularly targeted drug. This creates the opportunity to conduct targeted clinical trials with eligibility restricted to patients predicted to be responsive to the drug.**

Experimental Design: **We evaluated the relative efficiency of a targeted clinical trial design to an untargeted design for a randomized clinical trial comparing a new treatment to a control. Efficiency was evaluated with regard to number of patients required for randomization and number required for screening.**

Results: **The effectiveness of this design, relative to the more traditional design with broader eligibility, depends on multiple factors, including the proportion of responsive patients, the accuracy of the assay for predicting responsiveness, and the degree to which the mechanism of action of the drug is understood. Explicit formulas were derived for computing the relative efficiency of targeted *versus* untargeted designs.**

Conclusions: **Targeted clinical trials can dramatically reduce the number of patients required for study in cases where the mechanism of action of the drug is understood and an accurate assay for responsiveness is available.**

INTRODUCTION

Many cancer therapeutics benefit only a subset of treated patients. Genomic technologies such as DNA microarray expression profiling are providing biomarkers that facilitate the prediction of which patients are most likely to respond to a given regimen (1, 2). Molecularly targeted drugs are of increasing importance in cancer therapeutics, and such drugs are only expected to be effective for patients whose tumors express the target (3, 4). Thus, clinical trials may be increasingly tailored for patients who are predicted to respond to therapy (5). We call these targeted designs. As discussed in this article, we studied the efficiency of targeted designs in comparison with traditional randomized designs with broader eligibility criteria. We evaluated efficiency in the context of a binary outcome end point. Although many clinical trials use survival or time-to-progression end points, the binary end point setting is more tractable, and we obtained results that are intuitive and should be useful in understanding the factors that effect efficiency generally. For the untargeted and targeted design, we considered the comparison of a control *versus* experimental treatment with the same number of randomized patients in the two groups.

We compared the two designs with regard to the number of randomized patients required. We also compared the number of randomized patients for the untargeted design to the number of screened patients required for the targeted design. We assume that in the targeted design patients are screened using an assay that indicates whether the patient is likely to benefit from the new treatment. If the control arm is an active treatment, then the screening classifier should provide an indication of whether the patient is more likely to respond to the new regimen than to the control arm. Our efficiency comparisons are based on using the formula of Ury and Fleiss (6) for planning sample size for comparing proportions because of its known accuracy for approximating the tables of Casagrande, Pike, and Smith for the power of Fisher's exact test (7).

MATERIALS AND METHODS

We considered a population of patients consisting of an R+ portion who were predicted to be responsive to the new treatment and a remainder portion R−. The R− strata constituted a proportion γ of the population. Patients were randomized between the control and the experimental groups. p_c denotes the response probability in control group and was assumed to be the same for R− and R+ patients. The response probability in the treatment group was $p_c + \delta_0$ and $p_c + \delta_1$ for the R− and R+ patients, respectively. The response probability p_e for the experimental treatment group in the untargeted design was a weighted average of $p_c + \delta_0$ and $p_c + \delta_1$ with weights γ and 1-γ, respectively.

For the targeted design we added the symbol T. The response probability in the experimental group was $p_e^T = p_c + \delta_1$. We consider the one-sided test of the null hypothesis $p_c = p_e$ against the alternative hypothesis $p_e > p_c$.

Let n and n^T denote the number of patients needed to randomize in the untargeted and targeted design respectively to achieve the same statistical power for testing the null hypothesis. The expressions for n and n^T are indicated in the Appendix. The relative efficiency of the untargeted and the targeted designs can be expressed in the form:

$$n/n^T = \left[\frac{\delta_1}{\gamma\delta_0 + (1-\gamma)\delta_1}\right]^2 f \qquad \text{(A)}$$

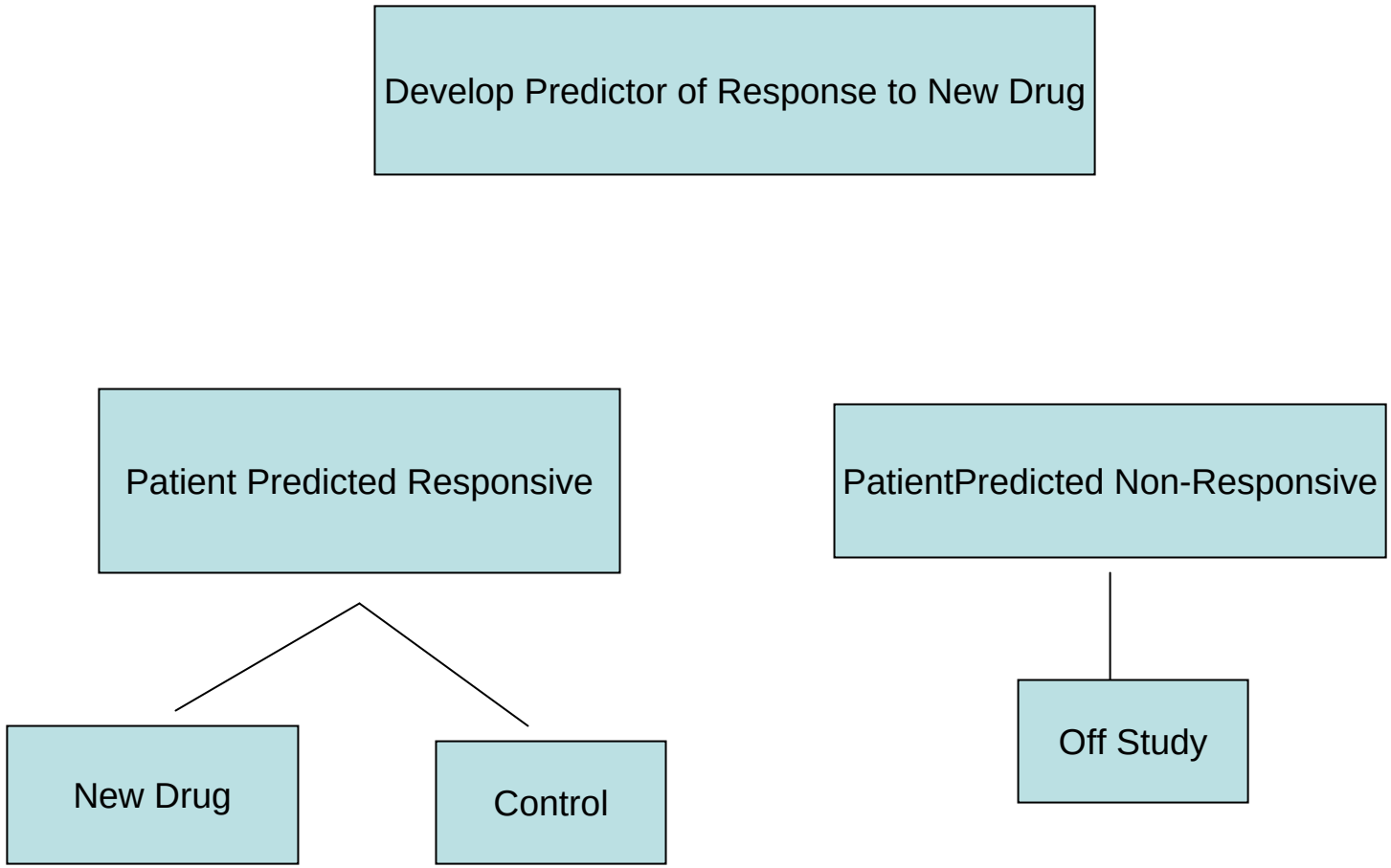
Received 3/11/04; revised 5/14/04; accepted 5/19/04.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
Note: Supplementary data for this article may be found at http://linus.nci.nih.gov/~brb/TechReport.htm.
Requests for reprints: Richard Simon, National Cancer Institute, 9000 Rockville Pike, Bethesda MD 20892-7434. Phone: 301-496-0975; Fax: 301-402-0560; E-mail: rsimon@nih.gov.

Pharmacogenomic Model for Two Treatments With Binary Response

- Molecularly targeted treatment E
- Control treatment C
- λ Proportion of patients predicted responsive (Assay+)
- p_c control response probability
- response probability for Assay+ patients receiving E is $(p_c + \delta_1)$
- Response probability for Assay- patients receiving E is $(p_c + \delta_0)$

Two Clinical Trial Designs

- Un-targeted design
  - Randomized comparison of E to C without screening for expression of molecular target
- Targeted design
  - Assay patients for expression of target
  - Randomize only patients expressing target


Develop Predictor of Response to New Drug
Patient Predicted Responsive
PatientPredicted Non-Responsive
New Drug
Control
Off Study

- Compare the two designs with regard to the number of patients required to achieve a fixed statistical power for detecting treatment effectiveness

Randomized Ratio
(normal approximation)

- RandRat = $n_{untargeted}/n_{targeted}$

$$RandRat \approx \left(\frac{\delta_1}{\lambda\delta_1 + (1-\lambda)\delta_0} \right)^2$$

- If δ_0=0, RandRat = $1/\lambda^2$
- If $\delta_0 = \delta_1/2$, RandRat = $4/(\lambda+1)^2$

Randomized Ratio $n_{untargeted}/n_{targeted}$

λ Assay+	δ_0=0	$\delta_0 = \delta_1/2$
0.75	1.78	1.31
0.5	4	1.78
0.25	16	2.56

Screened Ratio

- $N_{untargeted} = n_{untargeted}$
- $N_{targeted} = n_{targeted}/\lambda$
- $ScreenRat = N_{untargeted}/N_{targeted} = \lambda RandRat$

Screened Ratio

λ Assay+	δ_0=0	$\delta_0 = \delta_1/2$
0.75	1.33	0.98
0.5	2	0.89
0.25	4	0.64

STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:329–339
Published online 18 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1975

On the efficiency of targeted clinical trials

A. Maitournam and R. Simon*,†

Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892-7434, U.S.A.

SUMMARY

The development of genomics-based technologies is demonstrating that many common diseases are heterogeneous collections of molecularly distinct entities. Molecularly targeted therapeutics is often effective only for some subsets patients with a conventionally defined disease. We consider the problem of design of phase III randomized clinical trials for the evaluation of a molecularly targeted treatment when there is an assay predictive of which patients will be more responsive to the experimental treatment than to the control regimen. We compare the conventional randomized clinical trial design to a design based on randomizing only patients predicted to preferentially benefit from the new treatment. Trial designs are compared based on the required number of randomized patients and the expected number of patients screened for randomization eligibility. Relative efficiency depends upon the distribution of treatment effect across patient subsets, prevalence of the subset of patients who respond preferentially to the experimental treatment, and assay performance. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: genomics; clinical trials; molecularly targeted therapeutics; pharmacogenomics; sample size; normal mixture

1. INTRODUCTION

Patient responses to therapeutics are often heterogeneous. In oncology, for example, response rates of less than 50 per cent are not uncommon. Most drugs have potential side effects and hence the cost to the patient of receiving an ineffective drug can be substantial.

Genomic technologies such as DNA sequencing, mRNA transcript profiling, and comparative genomic hybridization [1] are providing evidence that many diseases are more molecularly heterogeneous than previously recognized. For example, substantial effort is currently placed in developing mutation signatures and gene expression signatures of tumors [2, 3]. Such studies provide insight into the heterogeneity of disease pathogenesis and enable molecular disease taxonomies to be defined. Some genetic profiling studies identify new therapeutic targets. In other cases, genomic profiling of disease tissue has provided accurate predictors of response

*Correspondence to: Richard Simon, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434, U.S.A.
†E-mail: rsimon@mail.nih.gov

Received September 2003
Accepted June 2004

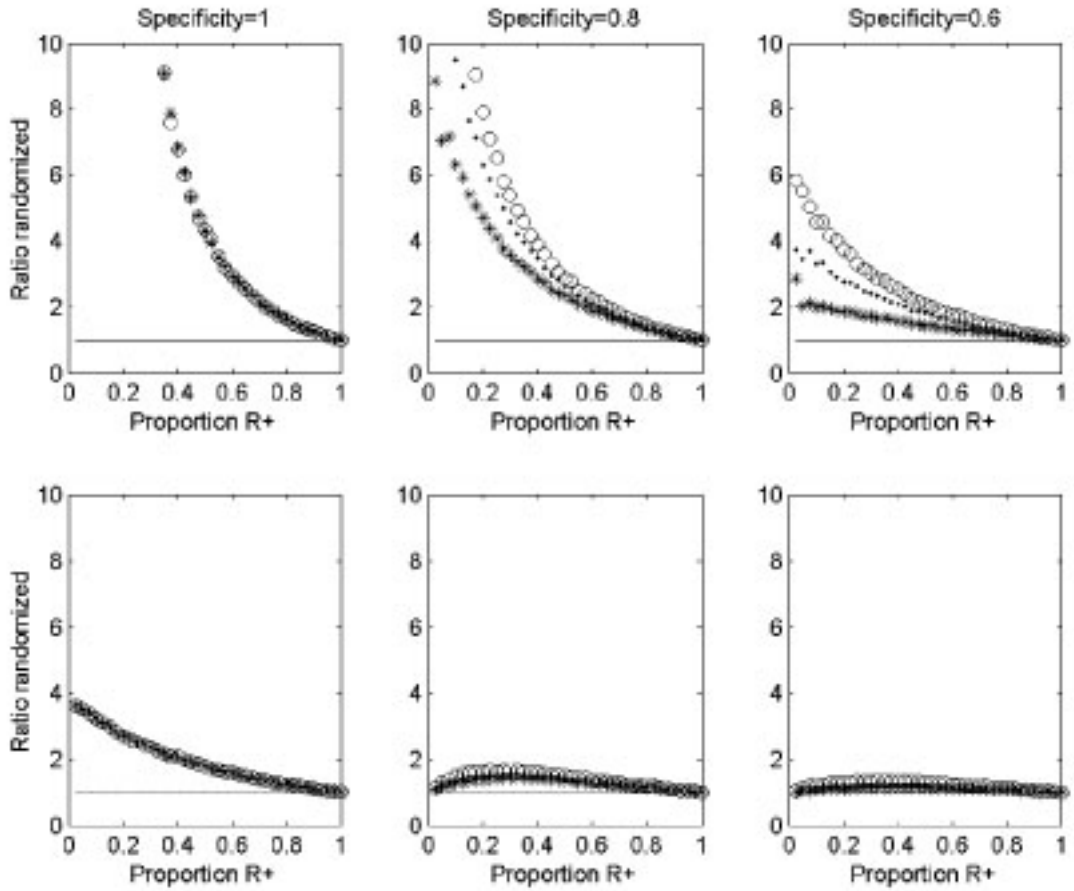


Figure 1. Ratio of number randomized for untargeted versus targeted designs. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ▫ Sensitivity = 1; • Sensitivity = 0.8; ⋆ Sensitivity = 0.6.

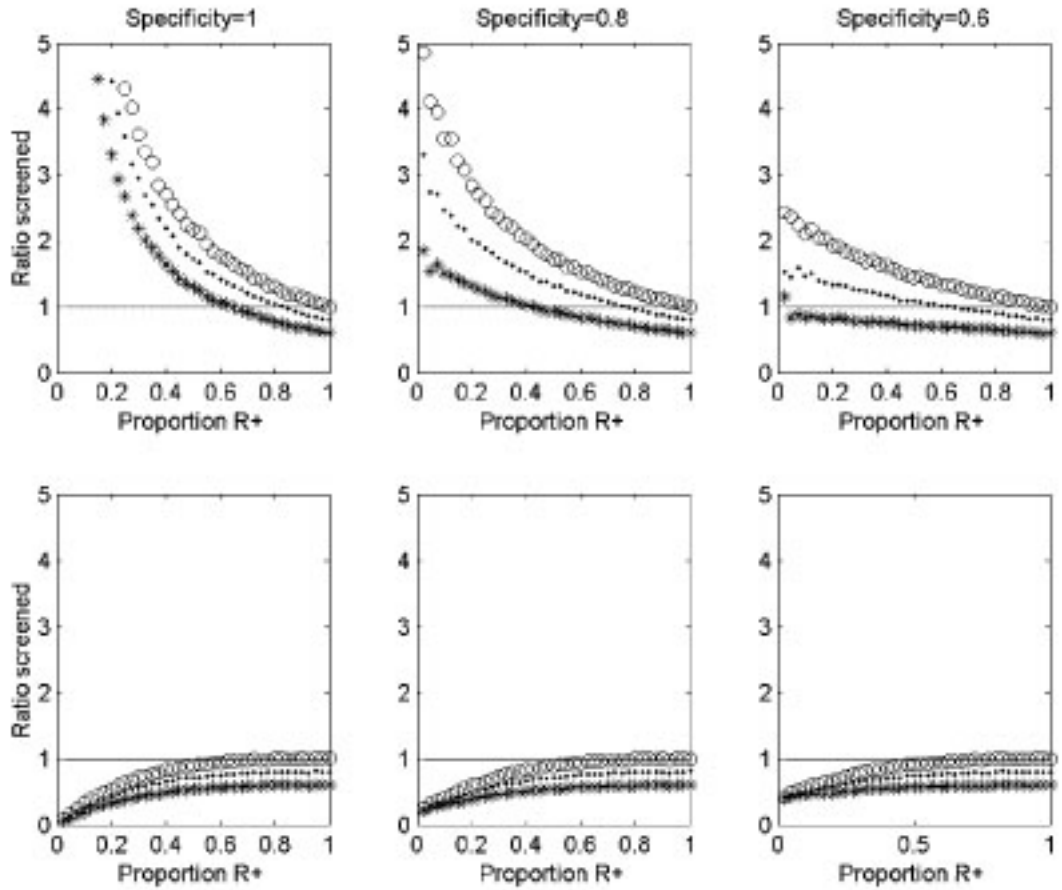


Figure 2. Ratio of number randomized for untargeted design to number screened for targeted design. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ∘ Sensitivity = 1; • Sensitivity = 0.8; ∗ Sensitivity = 0.6.

Adaptive Signature Design

An adaptive design for generating and prospectively testing a gene expression signature for sensitive patients

Boris Freidlin and Richard Simon

(Submitted for publication)

Adaptive Signature Design

- Randomized trial comparing E to C
  - Rapidly observed endpoint
- Stage 1 of accrual (half the patients)
  - Develop a binary classifier based on gene expression profile for the subset of patients that are predicted to preferentially benefit from the new treatment E compared to control C

Adaptive Signature Design
End of Trial Analysis

- Compare E to C for **all patients** at significance level 0.04
  - If overall H_0 is rejected, then claim effectiveness of E for eligible patients
  - Otherwise, compare E to C for patients accrued in second stage who are predicted responsive to E based on classifier developed during first stage.
    - Perform test at significance level 0.01
    - If H_0 is rejected, claim effectiveness of E for subset defined by classifier

Treatment effect restricted to subset.
10% of patients sensitive, 10 sensitivity genes, 10,000 genes, 400 patients.

Test	Power
Overall .05 level test	46.7
Overall .04 level test	43.1
Sensitive subset .01 level test (performed only when overall .04 level test is negative)	42.2
Overall adaptive signature design	85.3

Overall treatment effect, no subset effect.
10,000 genes, 400 patients.

Test	Power
Overall .05 level test	74.2
Overall .04 level test	70.9
Sensitive subset .01 level test	1.0
Overall adaptive signature design	70.9

Conclusions

- New technology and biological knowledge makes is increasingly feasible to identify which patients are most likely to benefit from a new treatment
- Targeting treatment can make it much easier to convincingly demonstrate treatment effectiveness
- Targeting treatment can greatly improve the therapeutic ratio of benefit to adverse effects, the proportion of treated patients who benefit

Conclusions

- Effectively defining and utilizing PG classifiers in drug development offers multiple challenges
- Much of the conventional wisdom about how to develop and utilize biomarkers is flawed and does not lead to definitive evidence of treatment benefit for a well defined population

Conclusions

- With careful prospective planning, genomic classifiers can be used in a manner that provides definitive evidence of treatment effect
  - Trial designs are available that will support broad labeling indications in cases where drug activity is sufficient, and the opportunity to obtain strong evidence of effectiveness in a well defined subset where overall effectiveness is not established

Conclusions

- Prospectively specified analysis plans for phase III data are essential to achieve reliable results
  - Biomarker analysis does not mean exploratory analysis except in developmental studies
  - Biomarker classifiers used in phase III evaluations should be completely specified based on external data
- In some cases, definitive evidence can be achieved from prospective analysis of patients in previously conducted clinical trials with extensive archival of pre-treatment specimens

Acknowledgements

- Boris Freidlin
- Aboubakar Maitournam
- Michael Radmacher
- Sudhir Varma
- Sue-Jane Wang