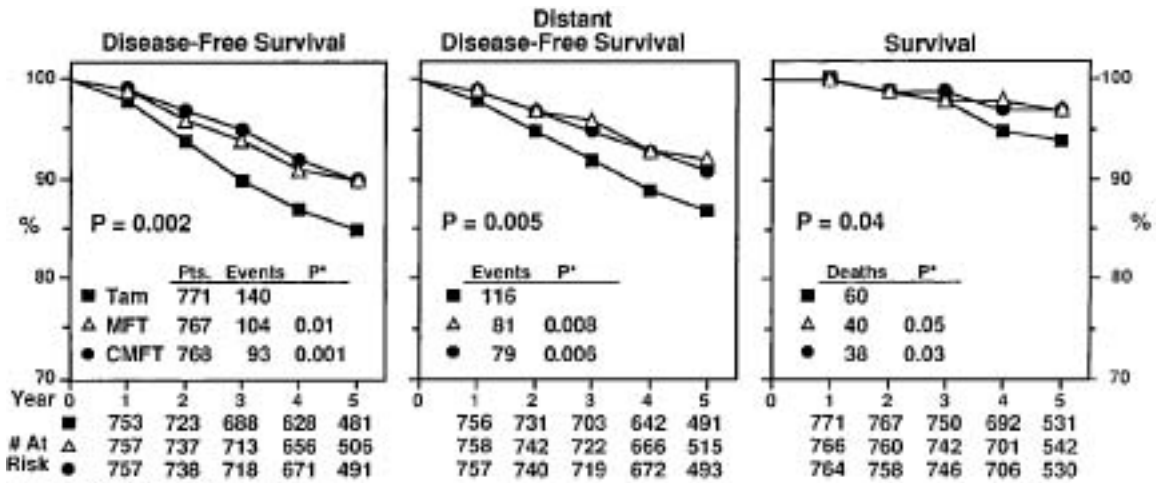
Design of Targeted Clinical Trials

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

BRB Website
http://linus.nci.nih.gov/brb

- Powerpoint presentations and audio files
- Reprints & Technical Reports
- BRB-ArrayTools software
- BRB-ArrayTools Data Archive
- Sample Size Planning for Targeted Clinical Trials

- Many cancer treatments benefit only a small proportion of the patients to which they are administered
- Targeting treatment to the right patients can greatly improve the therapeutic ratio of benefit to adverse effects
  - Smaller clinical trials needed
  - Treated patients benefit
  - Treatment more cost-effective for society



* Comparison to Tamoxifen

	RELATIVE RISK (95% CONFIDENCE INTERVAL)		
MFT/Tam	0.72 (0.56-0.93)	0.68 (0.51-0.90)	0.67 (0.45-0.99)
CMFT/Tam	0.65 (0.50-0.84)	0.67 (0.50-0.89)	0.64 (0.42-0.95)

- “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

Pharmacogenomic Targeting

- 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

“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

“Biomarkers”

- Surrogate endpoints
  - A measurement made before and after treatment to determine whether the treatment is working
  - Surrogate for clinical benefit
- Predictive classifiers
  - A measurement made before treatment to select good patient candidates for the treatment

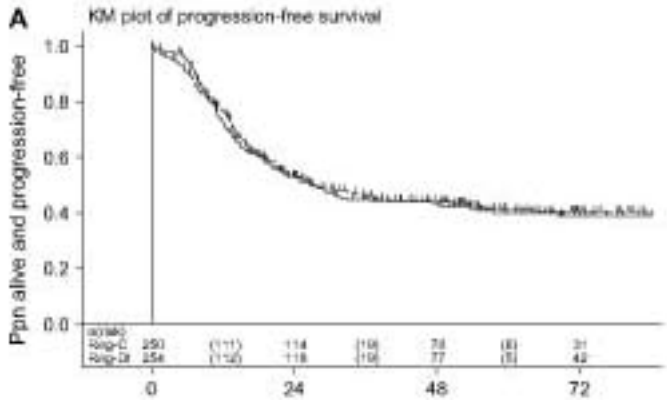
Surrogate Endpoints

- It is very difficult to properly validate a biomarker as a surrogate for clinical outcome. It requires a series of randomized trials with both the candidate biomarker and clinical outcome measured
  - Must demonstrate that treatment vs control differences for the candidate surrogate are concordant with the treatment vs control differences for clinical outcome
  - It is not sufficient to demonstrate that the biomarker responders survive longer than the biomarker non-responders

- “One rarely can establish that surrogate endpoints are valid. Even in that rare setting in which data on treatment Z would allow one to view S as a valid surrogate for T, one cannot extrapolate this surrogacy to any new treatment Z* that could have mechanisms of action that differ from those of Z.”
  - Fleming TR, Statistical Science 7:428-56, 1992

Cardiac Arrhythmia Supression Trial

- Ventricular premature beats was proposed as a surrogate for survival
- Antiarrythmic drugs supressed ventricular premature beats but killed patients at approximately 2.5 times that of placebo


A
KM plot of progression-free survival
Ppn alive and progression-free
1.0
0.8
0.6
0.4
0.2
0.0
0
24
48
72

- It is often more difficult and time consuming to properly “validate” an endpoint as a surrogate than to use the clinical endpoint in phase III trials
- The time frame for validating a surrogate is inconsistent with the time frame for initiating a pivotal study

- Biomarkers for use as endpoints in phase I or II studies need not be validated as surrogates for clinical outcome
- Unvalidated biomarkers can also be used for early “futility analyses” in phase III trials

Phase II/III Design

- Randomized trial comparing regimen containing new drug to control regimen
- Perform interim analysis comparing treatments using PFS (progression-free survival) endpoint
- If $p_{pfs}<p^*$ then continue trial to evaluate phase III endpoint
- Otherwise, terminate trial

Seamless Phase II/III Trial (a)

- Randomized comparison of standard treatment + new drug to standard treatment control
- Size trial using phase III (e.g. survival) endpoint
- Perform interim analysis using biomarker at a pre-specified time (e.g. after 50 patients per arm)
  - If treatment vs control results are not significant for biomarker, terminate accrual and do not claim any benefit of new treatment
  - If results are significant for biomarker, continue accrual and follow-up and do analysis of phase III endpoint at end of trial
    - Seek accelerated approval of new drug regimen based on significant biomarker result
- Interim analysis does not “consume” any of the significance level for the trial

Seamless Phase II/III Trial (b)

- Randomized comparison of 2 new drug regimens to control
- Size trial as phase III study with PFS endpoint
- Perform interim analysis using biomarker response
  - select new treatment arm with most promising biomarker response data
  - Continue accrual as 2-arm phase III trial of the selected treatment arm and the control arm
- Do analysis of phase III endpoint at end of trial using .025 level of significance
  - Or do permutation analysis to generate null distribution of test statistic resulting in less conservative analysis

Problems

- Surrogate endpoints
- Validity of biomarkers
- Hypothesis formulation and testing on the same set of data
- Conducting pivotal clinical trials without clearly pre-planned analysis
- “Stratification”

Validation=Fit for Purpose

- FDA terminology of “valid biomarker” and “probable valid biomarker” are not applicable to predictive classifiers
- “Validation” has meaning only as fitness for purpose and the purpose of predictive classifiers are completely different than for surrogate endpoints
- Criteria for validation of surrogate endpoints should not be applied to biomarkers used for treatment selection

- The components of multi-gene expression based classifiers should not have to be “valid biomarkers”
- It is often much easier to develop an accurate predictive classifier than to elucidate the role of the component genes in disease biology

Medicine Needs Predictive Markers not Prognostic Factors

- Most prognostic factors are not used because they are not therapeutically relevant
- Most prognostic factor studies are poorly designed and not focused on a clear objective; they use a convenience sample of patients for whom tissue is available. Generally the patients are too heterogeneous to support therapeutically relevant conclusions
- Prognostic and predictive studies should be designed with as much care and statistical rigor as clinical trials

Pusztai et al. The Oncologist 8:252-8, 2003

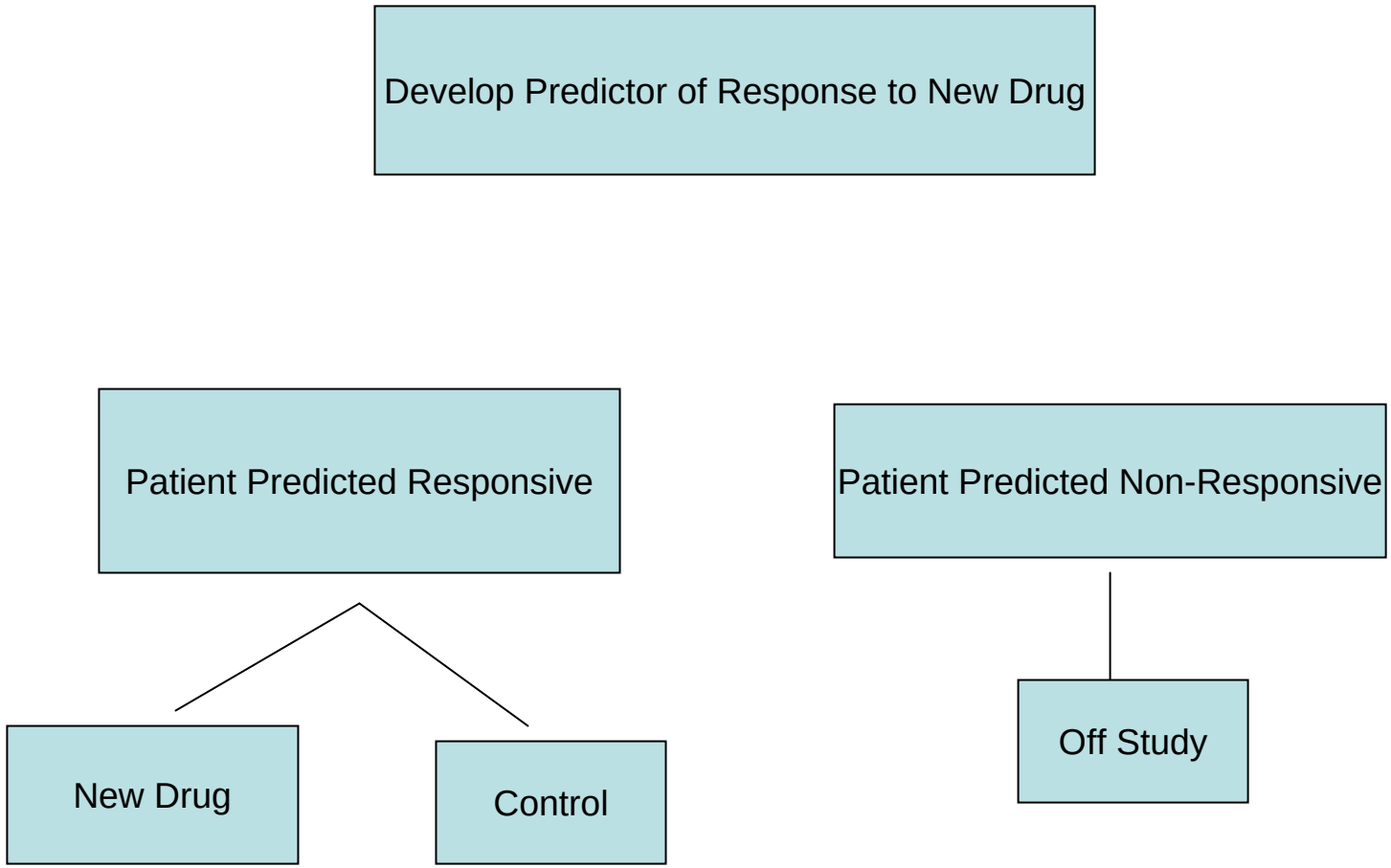
- 939 articles on “prognostic markers” or “prognostic factors” in breast cancer in past 20 years
- ASCO guidelines only recommend routine testing for ER, PR and HER-2 in breast cancer
- “With the exception of ER or progesterone receptor expression and HER-2 gene amplification, there are no clinically useful molecular predictors of response to any form of anticancer therapy.”

- Clinical trials of molecularly targeted drugs focused on patients whose tumors are expected to be susceptible to the drug can be much more efficient than traditional broad clinical trials

- In new drug development
  - The focus should be on evaluating the new drug in a population defined by a predictive classifier, not on “validating” the classifier
- In developing a predictive classifier for use in restricting a widely used treatment
  - The focus should be on evaluating the clinical utility of the classifier; Is clinical outcome better if the classifier is used than if it is not used?

New Drug Developmental Strategy (I)

- **Develop** a diagnostic classifier that identifies the patients likely to benefit from the new drug
- Develop a reproducible assay for the classifier
- **Use** the diagnostic to restrict eligibility to a prospectively planned evaluation of the new drug
- Demonstrate that the new drug is effective in the prospectively defined set of patients determined by the diagnostic


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

Applicability of Design I

- Primarily for settings where the classifier is based on a single gene whose protein product is the target of the drug
- With substantial biological basis for the classifier, it will often be unacceptable ethically to expose classifier negative patients to the new drug

Evaluating the Efficiency of Strategy (I)

- Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 10:6759-63, 2004.
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine 24:329-339, 2005.
- reprints and interactive sample size calculations at http://linus.nci.nih.gov/brb

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.

Two Clinical Trial Designs

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

Pharmacogenomic Model for Two Treatments With Binary Response

- Molecularly targeted treatment T
- Control treatment C
- 1-γ Proportion of patients that express target
  - Binary classifier
- p_c control response probability
- response probability for T patients who express target (R+) is ($p_c + \delta_1$)
- Response probability for T patients who do not express target (R-) is ($p_c + \delta_0$)

Untargeted Trial

- Compare outcome for treatment group T vs control group C without classifier data
- Fisher-Exact test at two-sided level .05 comparing response proportion in control group to response proportion in treatment group
- Number of responses in C group of n patients is binomial $B(n,p_c)$
- Number of responses in T group is
  - $B(n,(1-\gamma)(p_c+\delta_1)+ \gamma(p_c+\delta_0))$

- Determine n patients per treatment group for power $1-\beta$
  - Use Ury & Fleiss approximation Biom 36:347-51,1980.

Targeted Trial

- Compare outcome for treatment group T vs control group C for Assay positive patients
- Fisher-Exact test at two-sided level .05 comparing response proportion in control group to response proportion in treatment group
- Number of responses in C group of n patients is binomial $B(n,p_c)$
- Number of responses in T group is
  - $B(n,p_c+\delta_1)$

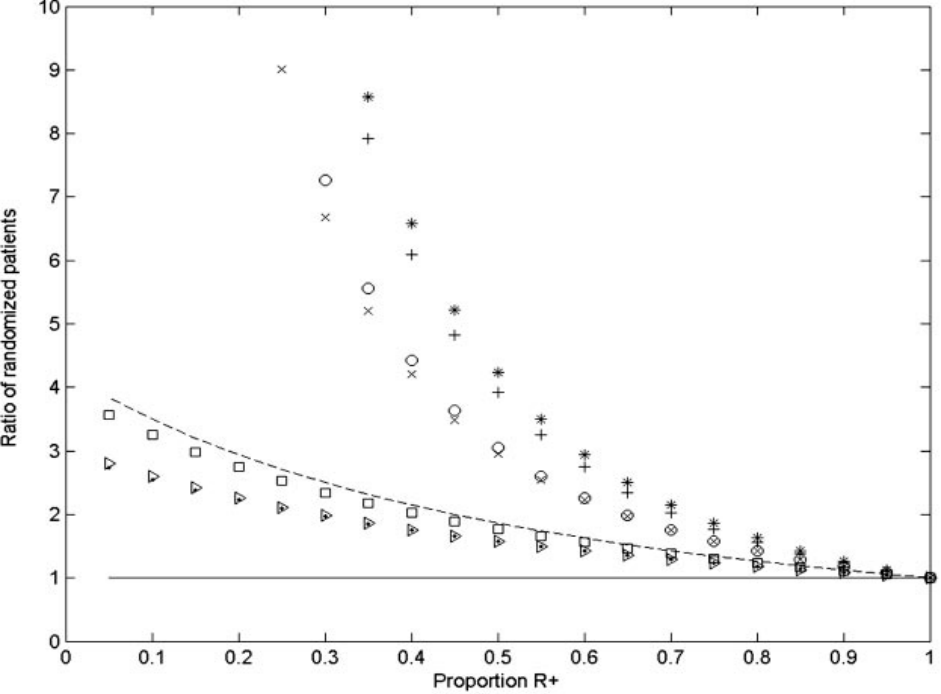
- Determine n^T patients per treatment group for power $1-\beta$
  - Use Ury & Fleiss approximation Biom 36:347-51,1980.

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

$$= \left[\frac{1}{1-\gamma} \right]^2 f \qquad \text{if } \delta_0 = 0$$

$$= \left[\frac{1}{1-\gamma/2} \right]^2 f \qquad \text{if } \delta_0 = \delta_1 / 2$$

- For p_c not close to 0
  - $f \sim 1$
- For p_c close to 0
  - $f < 1$


Ratio of randomized patients
Proportion R+
0
1
2
3
4
5
6
7
8
9
10
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

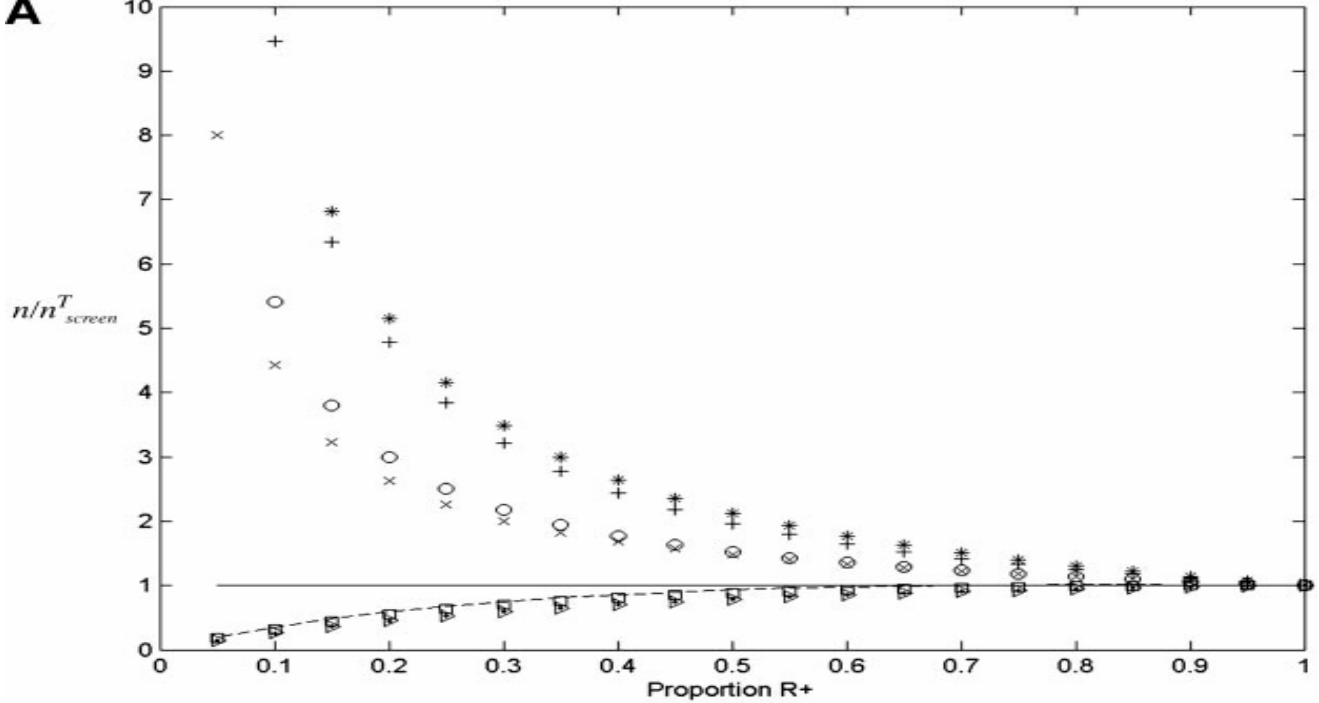
Screened Ratio

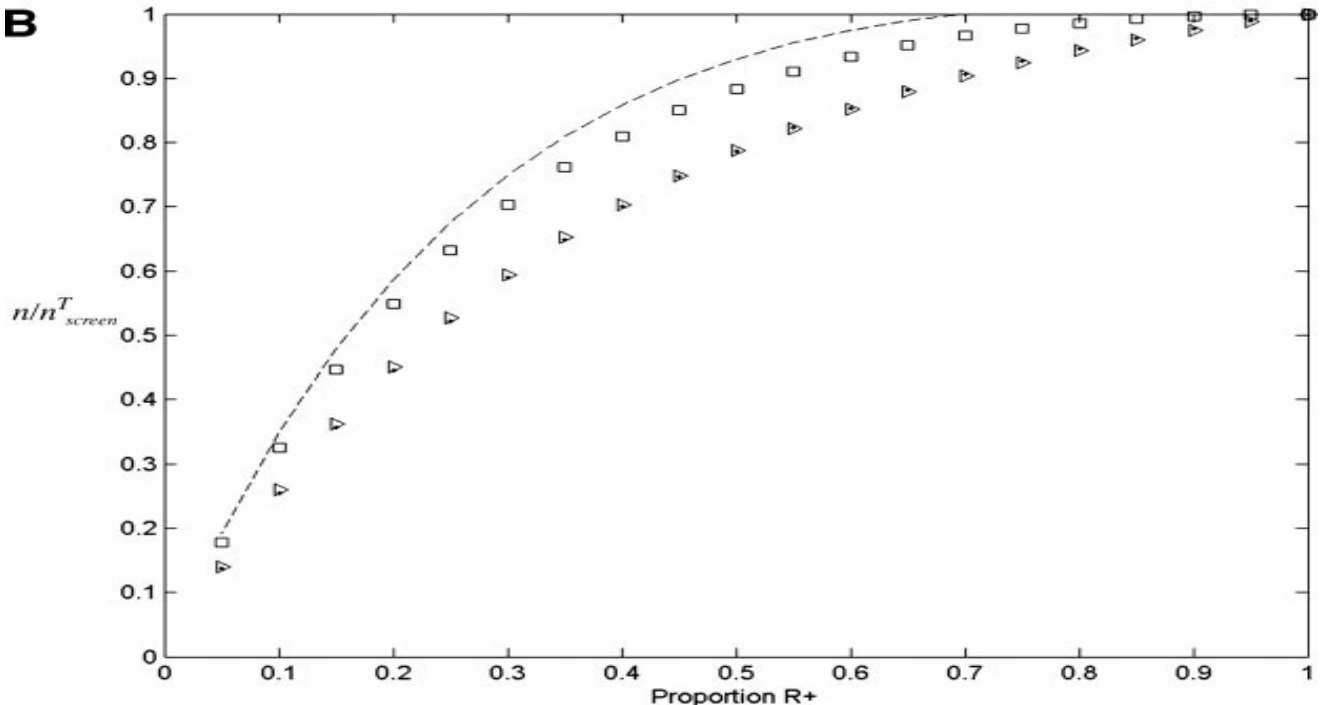
- $N_{untargeted} = n_{untargeted}$
- $N_{targeted} = n_{targeted}/(1-\gamma)$
- $ScreenRat = N_{untargeted}/N_{targeted}=(1-\gamma)RandRat$

$$n^T \approx (1-\gamma) n^T_{screen}$$

$$n / n^T_{screen} = \left[\frac{1}{1-\gamma} \right] f \quad \text{if } \delta_0 = 0$$

$$= \left[\frac{1-\gamma}{(1-\gamma/2)^2} \right] f \quad \text{if } \delta_0 = \delta_1 / 2$$


A
n/n^{T}_{screen}
Proportion R+


B
n/n^{T}_{screen}
Proportion R+

Imperfect Assay

- PPV = Pr(R+ | A+)
- NPV=Pr(R- | A-)

- Probability A+ patient responds to T
  - $p_c + PPV*\delta_1 + (1-PPV)* \delta_0$
- Prob R+ patient is excluded from targeted trial is
  - Pr(A- | R+) = 1-sensitivity
- Prob A- patient responds to T
  - $p_{c+}(1-NPV)* \delta_1 + NPV* \delta_0$

Approximations

For intuition but not made in published graphs

- Observed response rate ~ $N(p,p(1-p)/n)$
- $p_e(1-p_e) \sim p_c(1-p_c)$

Number of Randomized Patients Required

- Type I error α
- Power 1-β for obtaining significance

$$n = 2(p_c q_c + p_e q_e)\left(\frac{k_{1-\alpha} + k_{1-\beta}}{p_e - p_c}\right)^2$$

Randomized Ratio
(normal approximation)

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

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

- δ_1= rx effect in marker + patients
- δ_0= rx effect in marker - patients
- γ =proportion of marker - patients
- If δ_0=0, RandRat = $1/ (1-\gamma)^2$
- If $\delta_0 = \delta_1/2$, RandRat = $1/(1- \gamma/2)^2$

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

Proportion that Express target	$\delta_0=0$	$\delta_0=\delta_1/2$
0.75	1.78	1.31
0.5	4	1.78
0.25	16	2.56

Screened Ratio

Proportion that Express target	δ_0**=0**	δ_0**=** δ_1***/2***
0.75	1.33	0.98
0.5	2	0.89
0.25	4	0.64

Imperfect Assay Sensitivity & Specificity

- λ_{sens}=sensitivity
  - Pr[assay+ | target expressed]
- λ_{spec}=specificity
  - Pr[assay- | target not expressed]

Proportion of Assay Positive Patients That Express Target

$$PPV = \frac{(1-\gamma)\lambda_{sens}}{(1-\gamma)\lambda_{sens} + \gamma(1-\lambda_{spec})}$$

λ_{sens}	λ_{spec}	γ	PPV
0.9	0.9	0.75	0.96
0.9	0.9	0.5	0.9
0.9	0.9	0.25	0.75
0.9	0.9	0.10	0.50

Randomized Ratio

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

$$RandRat = \left(\frac{\text{effect in targeted trial}}{\text{effect in untargeted trial}} \right)^2$$

$$= \left(\frac{PPV\delta_1 + (1-PPV)\delta_0}{(1-\gamma)\delta_1 + \gamma\delta_0} \right)^2$$

Randomized Ratio sensitivity=specificity=0.9

Express target	δ_0=0	$\delta_0 = \delta_1/2$
0.75	1.29	1.26
0.5	1.8	1.6
0.25	3.0	1.96
0.1	25.0	1.86

Screened Ratio Imperfect Assay

- $N_{untargeted} = n_{untargeted}$

$$N_{\text{targeted}} \approx \frac{n_{\text{targeted}}}{(1-\gamma)\lambda_{sens} + \gamma(1-\lambda_{spec})}$$

$$\text{ScreenRat} = [(1-\gamma)\lambda_{sens} + \gamma(1-\lambda_{spec})]R\text{andrat}$$

Screened Ratio sensitivity=specificity=0.9

Express target	δ_0=0	$\delta_0 = \delta_1/2$
0.75	0.9	0.88
0.5	0.9	0.80
0.25	0.9	0.59
0.1	4.5	0.33

Trastuzumab

Herceptin

- Metastatic breast cancer
- 234 randomized patients per arm
- 90% power for 13.5% improvement in 1-year survival over 67% baseline at 2-sided .05 level
- If benefit were limited to the 25% assay + patients, overall improvement in survival would have been 3.375%
  - 4025 patients/arm would have been required
- If assay – patients benefited half as much, 627 patients per arm would have been required

Gefitinib

Iressa

- Two negative untargeted randomized trials first line advanced NSCLC
  - 2130 patients
- 10% have EGFR mutations
- If only mutation + patients benefit by 20% increase of 1-year survival, then 12,806 patients/arm are needed
- For trial targeted to patients with mutations, 138 are needed

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

Normal Outcome Model
Stat Med 24:329-39, 2005

- Control group outcome $X \sim N(0,1)$ for both types of patients
- Treatment group $Y \sim N(\delta_0,1)$ for R-
- Treatment group $Y \sim N(\delta_1,1)$ for R+
- Proportion R+ is $1-\gamma$

- For untargeted trial compare outcome for treatment group T vs control group C without classifier data
- Wilcoxon two-sample test at level .05 comparing X outcomes for control to Y outcomes for treatment group
  - Use mixture distribution of Y
- n patients per treatment group for power $1-\beta$

Power Approximately

$$1-\Phi\left\{\frac{0.5n^2+1.96\sqrt{n^2\frac{2n+1}{12}}-0.5-n^2p_1}{\mathrm{var}(W_{XY})}\right\}$$

$$\mathrm{var}(W_{XY})=n^2\left[p_1(1-p_1)+(n-1)(p_2+p_3-2p_1^2)\right]$$

$$p_1=\Pr(X<Y)$$

$$p_2=\Pr(X<Y_1,Y_2)$$

$$p_3=\Pr(X_1,X_2<Y)$$

- For targeted trial compare outcome for treatment group T vs control group C for Assay + patients
- Wilcoxon two-sample test at level .05 comparing X outcomes for control to Y outcomes for treatment group
  - Use distributions of Y for Assay + patients
- n_T patients per treatment group for power $1-\beta$

Relationship Between Assay & Receptor

- $Pr(A+|R+)=\lambda_{sens}$
- $Pr(A-|R-)=\lambda_{spec}$
- ω_{+}=PPV

$$\omega_{+} = \frac{(1-\gamma)\lambda_{sens}}{(1-\gamma)\lambda_{sens} + \gamma(1-\lambda_{spec})}$$

Effects of Assay Imprecision

- Assay imprecision doesn’t effect untargeted trial
- Control group outcomes in targeted trial are not effected by assay imprecision
- Treatment group in targeted trial is a mixture of R+ and R- patients because of imperfect specificity
- Targeted trial excludes some R+ patients because of imperfect sensitivity

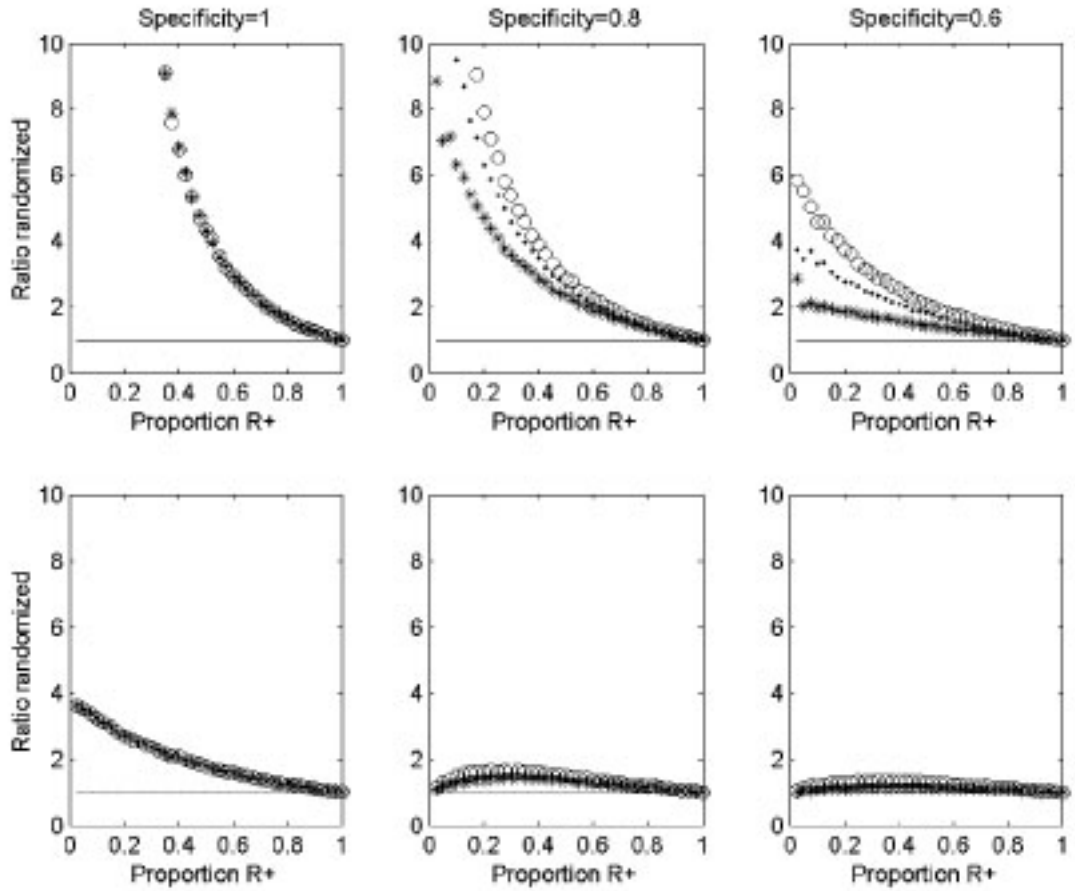


Figure 1. Ratio of number randomized for untargeted versus targeted designs. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ◦ Sensitivity = 1; • Sensitivity = 0.8; ⋆ Sensitivity = 0.6.

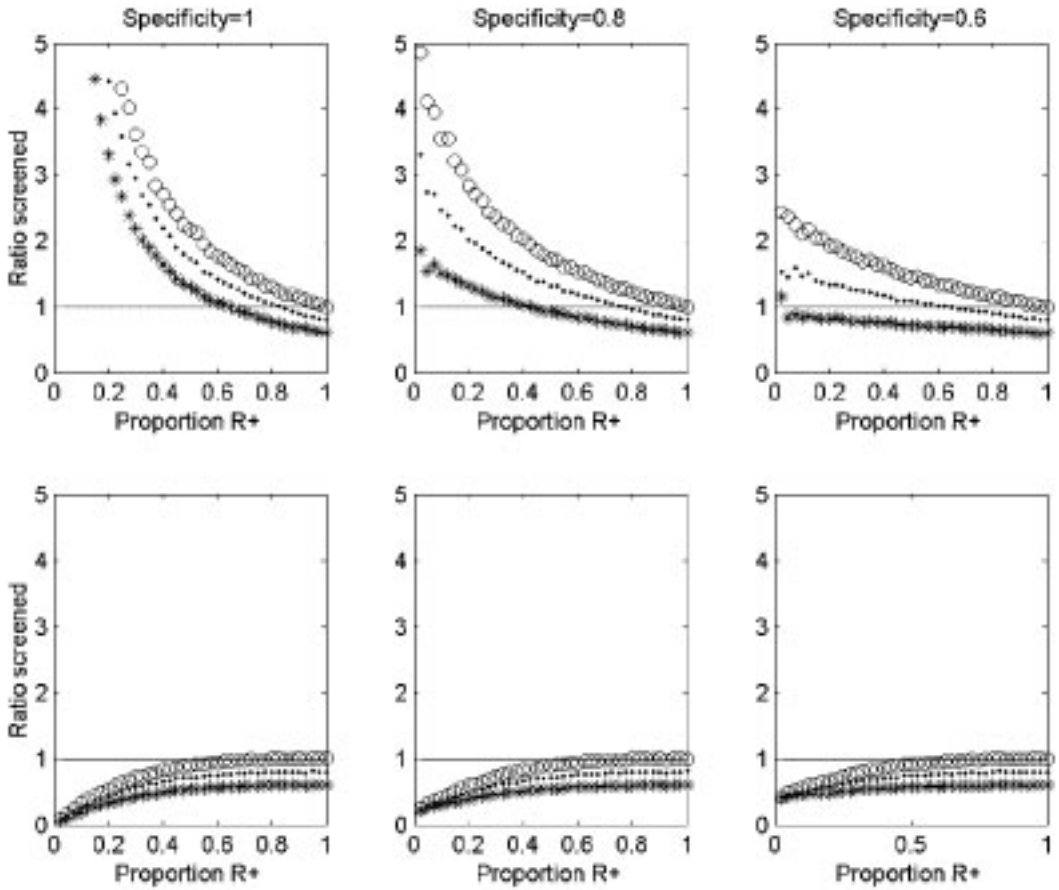


Figure 2. Ratio of number randomized for untargeted design to number screened for targeted design. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ∘ Sensitivity = 1; • Sensitivity = 0.8; ∗ Sensitivity = 0.6.

Comparison of Targeted to Untargeted Design

Simon R, Development and Validation of Biomarker Classifiers for Treatment Selection, JSPI

Treatment Hazard Ratio for Marker Positive Patients	Number of Events for Targeted Design	Number of Events for Traditional Design		
		Percent of Patients Marker Positive		
		20%	33%	50%
0.5	74	2040	720	316

Web Based Software for Comparing Sample Size Requirements

- http://linus.nci.nih.gov/brb/

biomathematics, and computational biology, on topics ranging from methodology to facilitate understanding at the molecular level of the pathogenesis of cancer to methodology to enhance the conduct of clinical trials of new therapeutic and diagnostic approaches.

Research Areas

Clinical trials, Drug Discovery, Molecular Cancer Diagnosis, Biomedical Imaging, Computational and Systems Biology, and Biostatistical Research

Technical Reports and Talks

Download the PDF version of our most recent research papers and PowerPoint version of the talk slides

BRB Staff

Investigators and contact information

BRB ArrayTools

Download the most advanced tools for microarray data analysis

BRB Alumni

Sample Size Calculation

BRB Annual Report 2005

Mathematics And Oncology

- The Norton-Simon Hypothesis
- The Norton-Simon Hypothesis and Breast Cancer Mortality in National Randomized Trial

Position Available

Post-doctoral fellow positions available

Software Download

- Accelerated Titration Design Software
- Optimal Two-Stage Phase II Design Software

Sample Size Calculation for Randomized Clinical Trials

- **Optimal Two-Stage Phase II Design**

- **Biomarker Targeted Randomized Design***

1. **Binary Outcome Endpoint**
2. **Survival and Time-to-Event Endpoint**

* Targeted design randomizes only marker positive patients to treatment or control arm. Untargeted design does not measure marker and randomizes all who otherwise are eligible.

Sample Size Calculation: Binary Outcome Endpoint

Evaluating the efficiency of targeted designs for randomized clinical trials and Supplement by Richard Simon and Aboubakar Maitournam. (Clinical Cancer Research 10:6759-6763, 2005)

pc	
gamma	
delta1	
delta0	
alpha	0.05
power	0.90

Submit

pc = probability of "response" for control arm

gamma = proportion of patients who are classifier negative (i.e. less responsive to new treatment

delta1 = improvement in response probability for new treatment in classifier positive patients

delta0 = improvement in response probability for new treatment in classifier negative patients

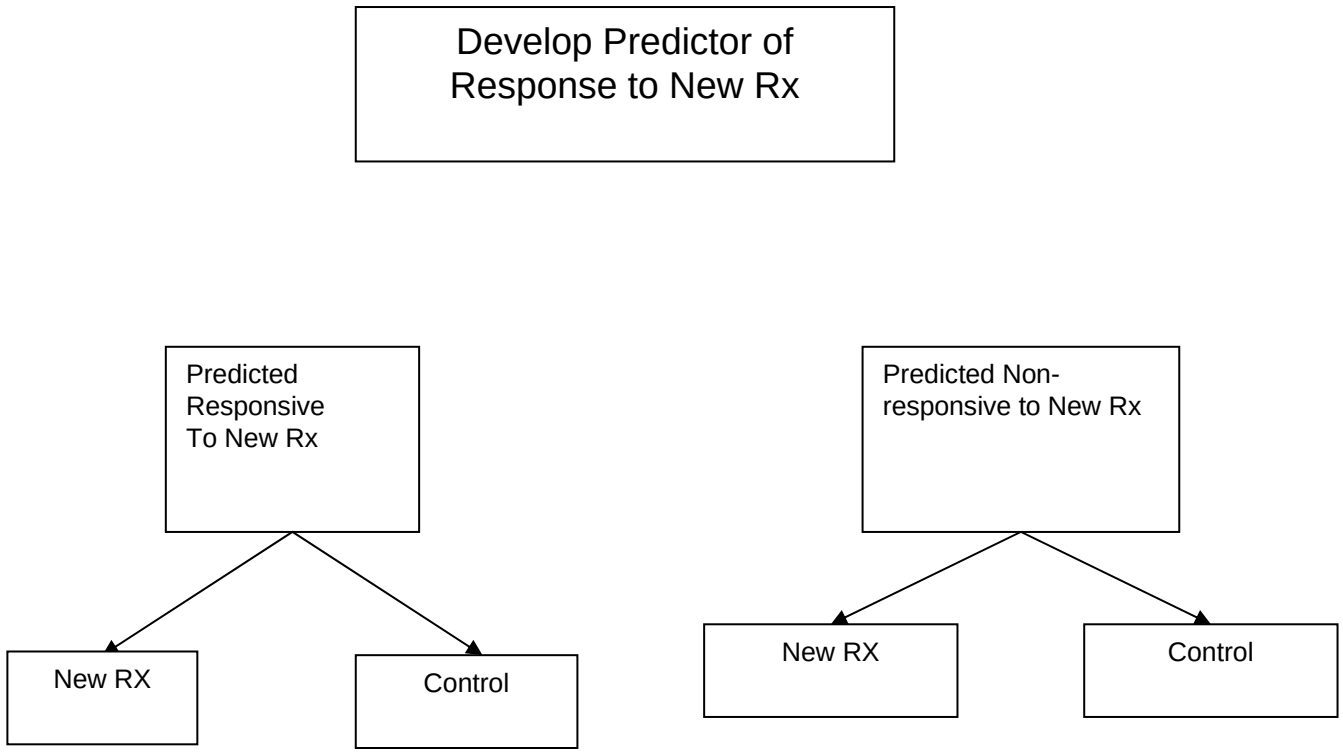
alpha = two-sided significance level

*Sample Size Calculation: Survival or Time-to-Event Endpoint**

	Median survival of the control group (years)
	or
	Proportion surviving beyond ____ years
	Total accrual rate (both marker positive and negative patients/year)
	Percent of patients marker negative
	% Reduction in hazard for treatment of marker positive patients
	% Reduction in hazard for treatment of marker negative patients
	Years of follow-up following end of accrual
0.05	Two-sided significance
0.90	Desired power for targeted design

Submit

Developmental Strategy (II)



Developmental Strategy (II)

- Do not use the diagnostic to restrict eligibility, but to structure a prospective analysis plan.
- Compare the new drug to the control overall for all patients ignoring the classifier.
  - If $p_{overall} \leq 0.04$ claim effectiveness for the eligible population as a whole
- Otherwise perform a single subset analysis evaluating the new drug in the classifier + patients
  - If $p_{subset} \leq 0.01$ claim effectiveness for the classifier + patients.

- This analysis strategy is designed to not penalize sponsors for having developed a classifier
- It provides sponsors with an incentive to develop genomic classifiers

Key Features of Design (II)

- The purpose of the RCT is to evaluate treatment T vs C overall and for the pre-defined subset; not to modify or refine the classifier or to re-evaluate the components of the classifier.
- This design assumes that the classifier is a binary classifier, not a “risk index”

Sample Size Planning for Design II

1. Size for standard power (e.g. 0.9) for detecting usual treatment effect at significance level 0.04
2. Size for standard power (e.g. 0.9) for detecting larger treatment effect in positive subset
3. Size as in 1 but extend accrual of classifier positive patients if overall test is non-significant

Hazard ratio δ to be detected	**Number of events required α=0.05**	**Number of events required α=0.04**
1.2	632	669
1.3	305	323
1.4	186	196
1.5	*128*	*135*
1.6	95	101
1.7	75	79
1.8	61	64
1.9	51	54
2.0	44	46

Number of events required for detecting a proportional hazard treatment effect with 90% power

Hazard ratio δ to be detected	**Number of events required α=0.01**
1.7	105
1.8	86
1.9	72
2.0	*62*
2.1	54
2.2	48
2.3	43

Number of events required for detecting a proportional hazard treatment effect with 90% power

Hazard rate to be detected overall	Hazard rate to be detected in + subset	Proportion classifier +	Number events needed for overall analysis at .04 level	Number events needed for classifier + analysis at .01 level	Number of total events to accrue
1.3	2	0.33	323	62	323
1.5	2	0.33	135	62	186

Patients per Arm

:

$$n \cong 2(p_c q_c + p_e q_e)\left\{\frac{k_{1-\alpha/2} + k_\beta}{p_e - p_c}\right\}^2$$

$$\cong 4 p_c q_c \left\{\frac{k_{1-\alpha/2} + k_\beta}{p_e - p_c}\right\}^2$$

Standard Design with 90% Power for Detecting Overall Treatment Effect Δ at Two-Sided Level 0.05

$$n_{.05} \cong 4 p_c q_c \left\{ \frac{1.96 + 1.28}{\Delta} \right\}^2$$

90% Power for Detecting Overall Treatment Effect Δ at Two-Sided Level 0.04

$$n_{.04} \cong 4 p_c q_c \left\{ \frac{2.054 + 1.28}{\Delta} \right\}^2$$

Ratio of Required Sample Size for .04 test to Required Sample Size for .05 test

$\{(2.054+1.28)/(1.96+1.28)\}^2 = 1.06$

Treatment Effect in the Classifier Positive Subset If Overall Effect is Δ

λ=1-γ

$$\lambda\delta_1 + (1-\lambda)\delta_0 = \Delta$$

$$\text{If } \delta_0 = 0 \text{ then } \delta_1 = \Delta / \lambda$$

$$\text{If } \delta_0 = \delta_1 / 2 \text{ then } \delta_1 = \Delta / \frac{1+\lambda}{2}$$

For 90% Power in the Classifier + Subset at 1% Significance

$$n_{.01} \cong 4 p_c q_c \left\{ \frac{2.576 + 1.28}{\Delta / \lambda} \right\}^2 \quad \text{if } \delta_0 = 0$$

$$n_{.01} \cong 4 p_c q_c \left\{ \frac{2.576 + 1.28}{\frac{\Delta}{\frac{1+\lambda}{2}}} \right\}^2 \quad \text{if } \delta_0 = \delta_1 / 2$$

Number of Total Patients Randomized per Treatment Needed to Have $n_{.01}$ per Treatment in Classifier + Subset

- $n_{.01}/\lambda$

If $\delta_0=0$

$$\frac{n_{.01}/\lambda}{n_{.05}} = \lambda\left\{\frac{2.576+1.28}{1.96+1.28}\right\}^2 = 1.41\lambda$$

If $\delta_0=0$

λ(fraction positive)	$\frac{n_{.01}/\lambda}{n_{.05}}$
0.5	0.71
0.25	0.35
0.10	0.14

If $\delta_0=\delta_1/2$

$$\frac{n_{.01}/\lambda}{n_{.05}} = 1.41\left\{\frac{\lambda+1}{2}\right\}^2 \frac{1}{\lambda}$$

If $\delta_0=\delta_1/2$

λ(fraction positive)	$\frac{n_{.01}/\lambda}{n_{.05}}$
0.5	1.59
0.25	2.20
0.10	4.27

FDA Subset Catch 22

- Do not accept claims based on subset analysis
- Require sponsors to do subset analysis to establish that a claim based on overall treatment effect applies to all subsets

Developmental Strategy III

- Do not use the diagnostic to restrict eligibility, but to structure a prospective analysis plan.
- Compare the new drug to the control for classifier positive patients
  - If $p_+ > 0.05$ make no claim of effectiveness
  - If $p_+ \leq 0.05$ claim effectiveness for the classifier positive patients and
    - Continue accrual of classifier negative patients and eventually test for smaller treatment effect at 0.05 level
    - Use sequential futility monitoring

Sample size Planning for III

- Accrue classifier positive and negative patients until there are sufficient classifier positive patients for standard power at significance level 0.05 for detecting larger than usual treatment effect D
- If treatment is found effective in classifier + patients, continue accrual of negative patients for standard power at significance level 0.05 for detecting smaller treatment effect d representing minimal useful clinical utility
  - Preform sequential futility monitoring to accrual of classifier - patients

Separate testing of treatment effect in positive and negative subsets

- With classifier tightly linked to drug target, it may be ethically unacceptable to expose classifier negative patients
- If it is ethically acceptable, some sponsors may prefer design III as it provides a primary focus on the classifier + patients
- With an empirically based classifier (C) in which there is limited confidence, design III will not be attractive as it requires commitment to a double sized clinical trial
  - The chance of a false negative in at least one subset is 19%
  - the potential value of being able to do a subset analysis may not be worth the cost of having to demonstrate effectiveness in both subsets separately for broad labeling

Predictive Medicine not Correlative Science

- The purpose of the RCT is to evaluate the new treatment overall and for the pre-defined subset
- The purpose is not to re-evaluate the components of the classifier, or to modify or refine the classifier
- The purpose is not to demonstrate that repeating the classifier development process on independent data results in the same classifier

The Roadmap

1. Develop a completely specified genomic classifier of the patients likely to benefit from a new drug
2. Establish reproducibility of measurement of the classifier
3. Use the completely specified classifier to design and analyze a new clinical trial to evaluate effectiveness of the new treatment with a pre-defined analysis plan.

Guiding Principle

- 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
    - And not closely regulated by FDA
  - Studies on which treatment effectiveness claims are to be based should be definitive studies that test a treatment hypothesis in a patient population completely pre-specified by the classifier

Use of Archived Samples

- From a non-targeted “negative” clinical trial to develop a binary classifier of a subset thought to benefit from treatment
- Test that subset hypothesis in a separate clinical trial
  - Prospective targeted type I trial
  - Prospective type II or III trial
  - Using archived specimens from a second previously conducted clinical trial

Development of Genomic Classifiers

- Single gene or protein based on knowledge of therapeutic target
- Empirically determined based on evaluation of a set of candidate classifiers
  - e.g. EGFR assays
- Empirically determined based on genome-wide correlating gene expression or genotype to patient outcome after treatment

Development of Genomic Classifiers

- During phase II development or
- After failed phase III trial using archived specimens.
- Adaptively during early portion of phase III trial.

Development of Genomic Classifiers

- Classifier for use with time-to-event endpoint
- Specificity of classifier for new treatment

Adaptive Signature Design

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

Boris Freidlin and Richard Simon

Clinical Cancer Research 11:7872-8, 2005

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

- Otherwise:
  - Using only the first half of patients accrued during the trial, develop a binary classifier that predicts the subset of patients most likely to benefit from the new treatment E compared to control C
  - Compare E to C for patients accrued in second stage who are predicted responsive to E based on classifier
    - Perform test at significance level 0.01
    - If H_0 is rejected, claim effectiveness of E for subset defined by classifier

True Model

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu^* + \lambda^* t_i + t_i \underline{\gamma}^* (x_{i1}, \ldots, x_{iK})$$

Classifier Development

- Using data from stage 1 patients, fit all single gene logistic models (j=1,…,M)
- Select genes with interaction significant at level η

$$\log it(p_i) = \mu + \lambda_j t_i + \nu_j x_{ij} + \beta_j t_i x_{ij}$$

Classification of Stage 2 Patients

- For i'th stage 2 patient, selected gene j votes to classify patient as preferentially sensitive to T if

$$\exp\left\{\hat{\lambda}_j + \hat{\beta}_j x_{ij}\right\} > R$$

Classification of Stage 2 Patients

- Classify i’th stage 2 patient as differentially sensitive to T relative to C if at least G selected genes vote for differential sensitivity of that patient

Simulation Parameters

- Gene expression levels of sensitivity genes MVN
  - mean m, variance v_1 and correlation r in sensitive patients
  - mean 0, variance v_2 and correlation r in non-sensitive patients
- Gene expression levels of other genes MVN with mean 0, variance v_0 and correlation r in all patients

- Treatment-expression interaction parameters (γ^*) same for all sensitivity genes
- γ^* value scaled (depending on K) so that log odds ratio of treatment effect is 5 for hypothetical patient with sensitivity gene expression levels at their expected values
  - i.e. m γ^*K=5
- Intercept μ scaled for control response rate of 25%

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

Treatment effect restricted to subset.
25% of patients sensitive, 10 sensitivity genes, 10,000 genes, 400 patients.

Test	Power
Overall .05 level test	99.0
Overall .04 level test	98.9
Sensitive subset .01 level test (performed only when overall .04 level test is negative)	99.7
Overall adaptive signature design	99.9

Overall treatment effect, no subset effect.
10% of patients sensitive, 10 sensitivity genes, 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

Stronger treatment effect for sensitive subset.
10% of patients sensitive, 10 sensitivity genes, 10,000 genes, 400 patients.

Test	Power
Overall .05 level test	97.0
Overall .04 level test	96.0
Sensitive subset .01 level test	45.6
Overall adaptive signature design	97.2

Empirical Power
RR for Control Patients 25%

Response Rate in Sensitive Subset	Overall .05	Overall .04	Subset .01	Overall Adaptive
98%	49.5	45.4	75.8	85.7
95%	43.0	38.5	63.1	75.0
87%	36.7	31.7	34.5	51.6
80%	31.6	28.4	17.6	38.8
71%	26.0	22.6	6.3	26.3

Biomarker Adaptive Threshold Design

Wenyu Jiang, Boris Freidlin & Richard Simon

(Submitted for publication)

http://linus.nci.nih.gov/brb

Biomarker Adaptive Threshold Design

- Randomized pivotal trial comparing new treatment E to control C
- Quantitative predictive biomarker B
- Survival or DFS endpoint

Biomarker Adaptive Threshold Design

- Have identified a univariate biomarker index B thought to be predictive of patients likely to benefit from E relative to C
- Eligibility not restricted by biomarker
- No threshold for biomarker determined
- Biomarker value scaled to range (0,1)

Procedure A

- Compare E vs C for all patients
  - If results are significant at level .04 claim broad effectiveness of E
  - Otherwise proceed as follows

Procedure A

- Test E vs C restricted to patients with biomarker B > b
  - Let S(b) be log likelihood ratio statistic
- Repeat for all values of b
- Let T = max{S(b)}
- Compute null distribution of T by permuting treatment labels
- If the data value of T is significant at 0.01 level, then claim effectiveness of E for a patient subset
- Compute point and interval estimates of the threshold b

Procedure B

- S(b)=log likelihood ratio statistic for treatment effect in subset of patients with $B \geq b$
- T=max{S(0)+R, max{S(b)}}
- Compute null distribution of T by permuting treatment labels
- If the data value of T is significant at 0.05 level, then reject null hypothesis that E is ineffective
- Compute point and interval estimates of the threshold b

Estimation of Threshold

$$\log h(t) = \log h_0(t) + \mu\tau + \eta I(B > b) + \gamma\tau I(B > b)$$

$\tau =$ binary treatment indicator

$l(\mu, \eta, \gamma, b) =$ log partial likelihood

$l(b) = \max\{l(\mu, \eta, \gamma, b)\}$

$\hat{b}$=argmax{l(b)}

$\hat{b}_* = \hat{b}$ value for bootstrap sample of cases

$\hat{F}_* =$ empirical distribution of $\hat{b}_*$

CI for b based on percentiles of $\hat{F}_*$

$\hat{F}_*(B) =$ probability patient with biomarker value B will benefit from treatment with E rather than C

Model	Hazard reduction for those who benefit	Overall Power	Adaptive Test
Everyone benefits	33%	.775	.751
50% benefit	60%	.888	.932
25% benefit	60%	.429	.604

Table 1. Empirical power: Procedures A and B vs. Overall Test

Sim #	Model	Effect size: reduction in hazard (hazard ratio)	Empirical Power		
			Overall test	Procedure A	Procedure B
1	Everybody benefits from new therapy	20% (.8)	.330	.304	.313
		33% (.67)	.775	.751	.732
		43% (.57)	.965	.957	.943
2	Only patients with biomarker values above 0.25 benefit from new therapy	43% (.57)	.819	.802	.837
		60% (.4)	.996	.997	.998
3	Only patients with biomarker values above 0.5 benefit from new therapy	43% (.57)	.505	.562	.607
		60% (.4)	.888	.932	.952
4	Only patients with biomarker values above 0.75 benefit from new therapy	43% (.57)	.196	.280	.311
		60% (.4)	.429	.604	.641
		69% (.31)	.600	.806	.846
5	Only patients with biomarker values above 0.9 benefit from new therapy	60% (.4)	.105	.238	.274
		69% (.31)	.162	.401	.412
		79% (.21)	.238	.632	.624
6	Linear increase in hazard ratio	43% (.57)*	.497	.504	.542
		60% (.4)*	.887	.892	.909
		69% (.31)*	.974	.981	.985
7	Linear increase in hazard ratio for patients with biomarker values above 0.5	43% (.57)**	.166	.212	.262
		60% (.4)**	.386	.514	.541
		69% (.31)**	.559	.744	.741

* maximum effect, effect increases linearly in biomarker from 0 to the maximum
** maximum effect, effect increases linearly in biomarker from .5 to the maximum

Prostate Cancer Data

DES (0.2 mg) vs Placebo

Covariate	# patients with measured covariate	Overall Test p value	Procedure A Stage 2 p value	Procedure B p value
AP	505	.084	.019	.041
SG	494	.110	.025	.050

Prostate Cancer Data

Covariate	# patients with measured covariate	Estimated Threshold	95% CI	80% CI
AP	505	36	(9,170)	(25,108)
SG	494	11	(10,13)	(11,11)

Figure 1: Estimated Probability of Benefit for AP

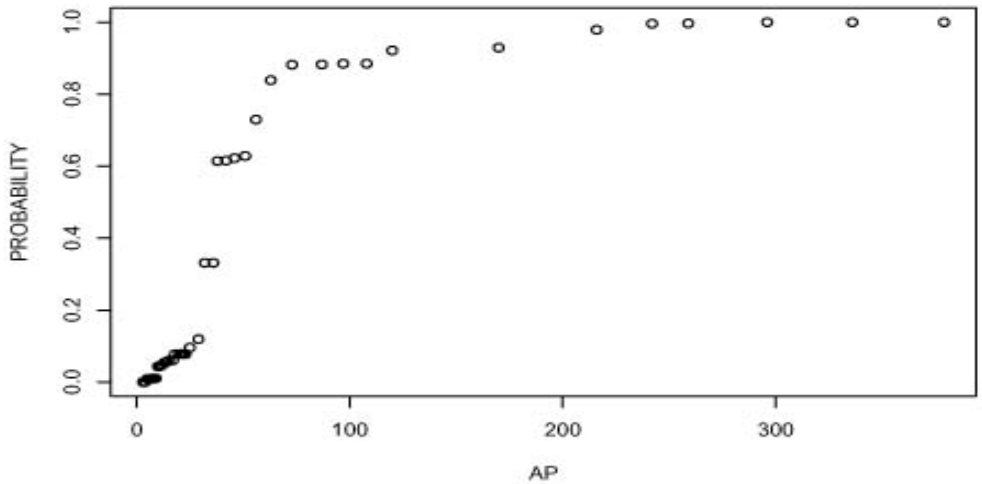
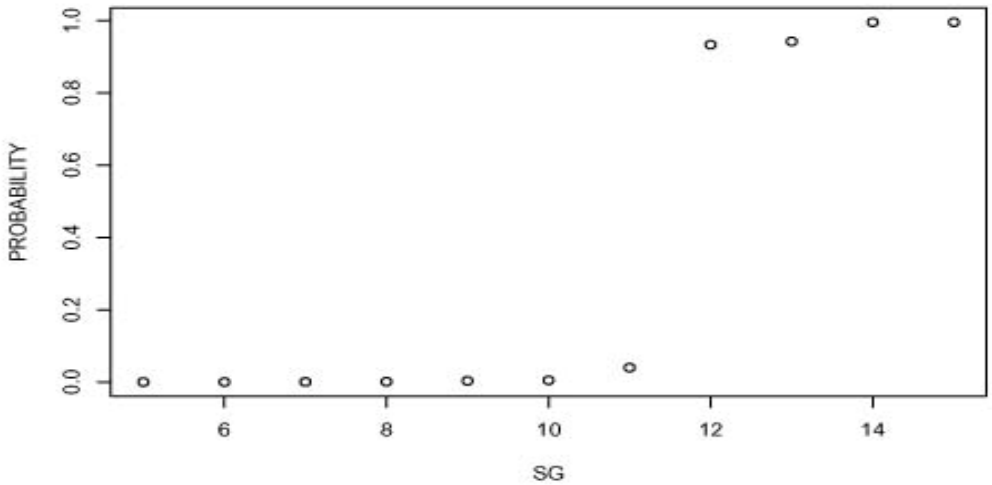


Figure 2: Estimated Probability of Benefit for SG



Sample Size Planning (A)

- Standard broad eligibility trial were designed for 80% power to detect reduction in hazard D at significance level 5%
- Biomarker adaptive design is sized for 80% power to detect same reduction in hazard D at significance level 4% for overall analysis

Estimated Power of Broad Eligibility Design (n=386 events) vs Adaptive Design A (n=412 events) 80% power for 30% hazard reduction

Model	Broad Eligibility Design	Biomarker Adaptive Design A
40% reduction in 50% of patients (22% overall reduction)	.70	.78
60% reduction in 25% of patients (20% overall reduction)	.65	.91
79% reduction in 10% of patients (14% overall reduction)	.35	.93

Sample Size Planning (B)

- Estimate power of procedure B relative to standard broad eligibility trial based on Table 1 for the row corresponding to the expected proportion of sensitive patients (π) and the target hazard ratio for sensitive patients
  - e.g. π=25% and Δ=.4 gives RE=.429/.641=.67
- When B has power 80%, overall test has power 80*.67=53%
- Use formula B.2 to determine the approximate number of events needed for overall test to have power 53% for detecting Δ=.4 limited to π=25% of patients

Events needed to Detect Hazard Ratio Δ With Proportional Hazards

$$D = 4\left(\frac{z_{1-\alpha} + z_{1-\beta}}{\log \Delta} \right)^2$$

Events (D') Needed for Overall Test to Detect Hazard Ratio Δ Limited to Fraction π

$$D' \approx D / \pi^2$$

Example Sample Size Planning for Procedure B

- Design a trial to detect Δ=0.4 (60% reduction) limited to π=25% of patients
  - Relative efficiency from Table 1 .429/.641=.67
- When procedure B has power 80%, standard test has power 80%*.67=53%
- Formula B.2 gives D'=230 events to have 53% power for overall test and thus approximate 80% power for B
- Overall test needs D=472 events for 80% power for detecting the diluted treatment effect

Table 5 The approximate number of events required to detect a 60% reduction in hazard in a subset of patients using broad eligibility design vs. biomarker adaptive design B

Model	Number of events	
	Broad eligibility design*	Biomarker adaptive design B**
Only patients with biomarker values above 0.25 benefit from new therapy (75% of patients are sensitive)	54	54
Only patients with biomarker values above 0.5 benefit from new therapy (50% of patients are sensitive)	118	102
Only patients with biomarker values above 0.75 benefit from new therapy (25% of patients are sensitive)	472	230
Only patients with biomarker values above 0.9 benefit from new therapy (10% of patients are sensitive)	2946	618

* numbers are for 80% power

** numbers are based on relative efficiency form Table 1 for hazard ratio of .4

Developing Gene Expression Based Classifiers

ARTICLE

Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting

Alain Dupuy, Richard M. Simon

Background Both the validity and the reproducibility of microarray-based clinical research have been challenged. There is a need for critical review of the statistical analysis and reporting in published microarray studies that focus on cancer-related clinical outcomes.

Methods Studies published through 2004 in which microarray-based gene expression profiles were analyzed for their relation to a clinical cancer outcome were identified through a Medline search followed by hand screening of abstracts and full text articles. Studies that were eligible for our analysis addressed one or more outcomes that were either an event occurring during follow-up, such as death or relapse, or a therapeutic response. We recorded descriptive characteristics for all the selected studies. A critical review of outcome-related statistical analyses was undertaken for the articles published in 2004.

Results Ninety studies were identified, and their descriptive characteristics are presented. Sixty-eight (76%) were published in journals of impact factor greater than 6. A detailed account of the 42 studies (47%) published in 2004 is reported. Twenty-one (50%) of them contained at least one of the following three basic flaws: 1) in outcome-related gene finding, an unstated, unclear, or inadequate control for multiple testing; 2) in class discovery, a spurious claim of correlation between clusters and clinical outcome, made after clustering samples using a selection of outcome-related differentially expressed genes; or 3) in supervised prediction, a biased estimation of the prediction accuracy through an incorrect cross-validation procedure.

Conclusions The most common and serious mistakes and misunderstandings recorded in published studies are described and illustrated. Based on this analysis, a proposal of guidelines for statistical analysis and reporting for clinical microarray studies, presented as a checklist of "Do's and Don'ts," is provided.

J Natl Cancer Inst 2007;99:147–57

DNA microarray technology has found many applications in biomedical research. In oncology, it is being used to better understand the biological mechanisms underlying oncogenesis, to discover new targets and new drugs, and to develop classifiers (predictors of good outcome versus poor outcome) for tailoring individualized treatments (1–4). Microarray-based clinical research is a recent and active area, with an exponentially growing number of publications. Both the reproducibility and validity of findings have been challenged, however (5,6). In our experience, microarray-based clinical investigations have generated both unrealistic hype and excessive skepticism. We reviewed published microarray studies in which gene expression data are analyzed for relationships with cancer outcomes, and we propose guidelines for statistical analysis and reporting, based on the most common and serious problems identified.

Medicine, followed by hand screening of abstracts and articles. The detailed process of selection is presented in Supplementary Note 1 (available online). The inclusion criteria were as follows: the work was an original clinical study on human cancer patients, published in English before December 31, 2004; it analyzed gene expression data of more than 1000 spots; and it presented statistical analyses relating the gene expression profiling to a clinical outcome. Two types of outcome were considered: 1) A relapse or death occurring during the course of the disease. 2) A therapeutic response.

Affiliations of authors: Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD (AD, RMS); Université Paris VII Denis Diderot, Paris, France (AD); Assistance Publique-Hôpitaux de Paris, Service de Dermatologie, Hôpital Saint-Louis, Paris, France (AD).

Correspondence to: Richard M. Simon, DSc, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892 (e-mail: rsimon@nih.gov).

Major Flaws Found in 40 Studies Published in 2004

- Misleading use of cluster analysis
  - 13/28 studies invalidly claimed that expression clusters based on differentially expressed genes could help distinguish clinical outcomes
- Inadequate control of multiple comparisons in gene finding
  - 9/23 studies had unclear or inadequate methods to deal with false positives
    - 10,000 genes x .05 significance level = 500 false positives
- Misleading report of prediction accuracy
  - 12/28 reports based on incomplete cross-validation
- 50% of studies contained one or more major flaws

and criteria used for selection of cases.

9	Don't	Transform time-to-outcome data into a binary outcome variable if the goal is to predict groups with different survival probabilities.	Use statistical methods suited for time-to-event data, unless you can ensure the absence of bias due to transformation. See text and Supplementary Fig. 2 (available online).
Outcome-related gene finding†			
10	Don't	Use only fold changes between groups to select the differentially expressed genes.	This does not take into account the variance of the genes' data values.
11	Don't	Use a .05 *P* value threshold to select the differentially expressed genes.	A set of 10000 genes will yield on average 500 false-positive genes if this threshold is used.
12	Do	Use a method for controlling the number of falsely differentially expressed genes.	Lowering the *P* value threshold for selection (e.g., to .001) is the simplest method. Others are available.
13	Do	Use a permutation test to assess the probability of finding the same number of differentially expressed genes as the one you found from your dataset.	The result should be significant at .05 *P* value level.
Class discovery			
14	Don't	Use class discovery methods if you are interested in classifying new samples in the future.	Supervised prediction should be used for this purpose. It utilizes the outcome information to optimize predictive accuracy. See text.
15	Don't	Use a selection of outcome-related differentially expressed genes if you intend to correlate cluster-defined classes with the outcome.	Supervised clustering leads to a spurious correlation between cluster and outcome. See text and Fig. 1.
16	Don't	Select the clustering method that gives the best result.	Class discovery should not be result driven.
17	Do	Use methods for testing the reproducibility of cluster finding.	Assessing the reproducibility of cluster finding without using external information makes class discovery more convincing. See text.
18	Don't	Use conventional statistical tests for computing the statistical significance of genes that are differentially expressed between two clusters.	These tests assume independence between class definition and expression profile data, which is not the case for cluster-defined classes.
Supervised prediction			
19	Do	Frame a therapeutically relevant question and select a homogeneous set of patients accordingly.	Classifiers developed outside a specific therapeutically relevant context are unlikely to be useful and utilized. See text.
20	Don't	Violate the fundamental principle of classifier validation, i.e., no preliminary use of the tested samples.	Most of the "Don't" items on validation procedures are illustrations of how this principle can be violated. See text and Fig. 2 and Supplementary Fig.1 (available online).
21	Don't	Attempt to predict cluster-defined classes.	Classes should be defined independently from the expression profile data.
Evaluating the prediction on a separate test set			
22	Don't	Use any information from the test set for developing the classifier.	The test set is to be used exclusively for evaluating the classifier performance. See text and Fig. 2.
23	Do	Access the test set only once and only for testing the samples with the fully specified classifier developed from the training set.	The test set must not be used to choose the best classifier. See text and Fig. 2.
24	Do	Use the same outcome definition as the one used in the training set.	

(Table continues)

Checklist			Comment
	Evaluating the prediction with a cross-validation procedure		
25	Don't	Use all the samples from the dataset to develop the classifier and test them.	The resubstitution estimate is not a cross-validation procedure. See text and Fig. 2.
26	Don't	Use the same feature selection for all iterations.	This inflates the estimate of the prediction accuracy. See text and Fig. 2.
27	Don't	Perform a cross-validation procedure on a selection of outcome-related differentially expressed genes.	Idem. Invalid although commonly done.
28	Do	Report the estimates for all the classification algorithms if several have been tested, not just the most accurate.	
29	Don't	Consider that testing a few additional independent samples adds value to a correctly cross-validated estimate of the classifier prediction accuracy.	However, this may be valuable if the additional samples are in sufficient number and are representative of the samples in which the classifier might be used in the future. See text.
30	Do	Report the fully specified classifier with its parameters.	So it can be used by others. Parameters are obtained from the whole training set in a separate test set procedure and from the whole dataset in a cross-validation procedure.
31	Do	Report the correctly validated sensitivity and specificity or positive and negative apparent predictive values (for a binary outcome).	Receiver-operating characteristic curves may also be used. See text.
32	Don't	Use an odds ratio to assess the performance of the prediction (for a binary outcome).	The odds ratio is a measure of association, not of prediction accuracy. See text and Supplementary Fig. 3 (available online).
33	Do	Report the statistical significance of the prediction accuracy and, even better, of the sensitivity and specificity (for a binary outcome).	It states the probability of obtaining a prediction accuracy as high as actually observed if there was no relationship between the expression data and the outcome. See text.
34	Don't	Use a Fisher's exact test or chi-square test to assess the statistical significance of the prediction accuracy for a binary outcome.	They do not test the statistical significance of the prediction. See text and Supplementary Fig. 3 (available online).
35	Do	Pay attention to the imbalance between outcome categories when interpreting the prediction accuracy of a binary outcome.	90% prediction accuracy may be inadequate if outcome categories are highly imbalanced. See text and Supplementary Fig. 3 (available online).
36	Don't	Use the log-rank test for testing the difference in survival between cross-validated groups.	The test is invalid because of a dependency among cases after cross-validation.
37	Don't	Use standard regression models, e.g., logistic regression or proportional hazards model, with cross-validated predicted groups.	Idem.
38	Don't	Assess the utility of the prediction based on the value of the regression coefficient or on its *P* value from multivariable regression models.	Regression coefficients are poor measures of prediction accuracy, and the test of statistical significance simply assesses if the coefficient is different from 0. See text.
39	Do	Assess the added value of the classifier by examining its performance within the levels of the standard prognostic factors.	Other approaches can be used. See text.
40	Do	Assess the utility of the classifier in a clinical context, for the therapeutically relevant question, and plan, if appropriate, further studies for external validation.	

Good Microarray Studies Have Clear Objectives

- Class Comparison
  - Find genes whose expression differs among predetermined classes, e.g. tissue or experimental condition
- Class Prediction
  - Prediction of predetermined class (e.g. treatment outcome) using information from gene expression profile
- Class Discovery
  - Discover clusters of specimens having similar expression profiles
  - Discover clusters of genes having similar expression profiles

Class Comparison and Class Prediction

- Not clustering problems
- Supervised methods

Class Prediction

- A set of genes is not a classifier
- Testing whether analysis of independent data results in selection of the same set of genes is not an appropriate test of predictive accuracy of a classifier

Myth

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

- Artificial intelligence sells to non-specialists who cannot distinguish hype from substance.
- Comparative studies generally indicate that simpler methods work as well or better for microarray problems because they avoid over-fitting the data.

Linear Classifiers for Two Classes

- Fisher linear discriminant analysis
- Diagonal linear discriminant analysis (DLDA) assumes features are uncorrelated
- Compound covariate predictor (Radmacher et al)
- Weighted voting (Golub et al.)
- Support vector machines with inner product kernel
- Perceptron (Khan et al.)

Other Simple Methods

- Nearest neighbor classification
- Nearest k-neighbors
- Nearest centroid classification
- Shrunken centroid classification

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

ORIGINAL ARTICLE

Concordance among Gene-Expression–Based Predictors for Breast Cancer

Cheng Fan, M.S., Daniel S. Oh, Ph.D., Lodewyk Wessels, Ph.D., Britta Weigelt, Ph.D., Dimitry S.A. Nuyten, M.D., Andrew B. Nobel, Ph.D., Laura J. van't Veer, Ph.D., and Charles M. Perou, Ph.D.

ABSTRACT

BACKGROUND

Gene-expression–profiling studies of primary breast tumors performed by different laboratories have resulted in the identification of a number of distinct prognostic profiles, or gene sets, with little overlap in terms of gene identity.

METHODS

To compare the predictions derived from these gene sets for individual samples, we obtained a single data set of 295 samples and applied five gene-expression–based models: intrinsic subtypes, 70-gene profile, wound response, recurrence score, and the two-gene ratio (for patients who had been treated with tamoxifen).

RESULTS

We found that most models had high rates of concordance in their outcome predictions for the individual samples. In particular, almost all tumors identified as having an intrinsic subtype of basal-like, HER2-positive and estrogen-receptor–negative, or luminal B (associated with a poor prognosis) were also classified as having a poor 70-gene profile, activated wound response, and high recurrence score. The 70-gene and recurrence-score models, which are beginning to be used in the clinical setting, showed 77 to 81 percent agreement in outcome classification.

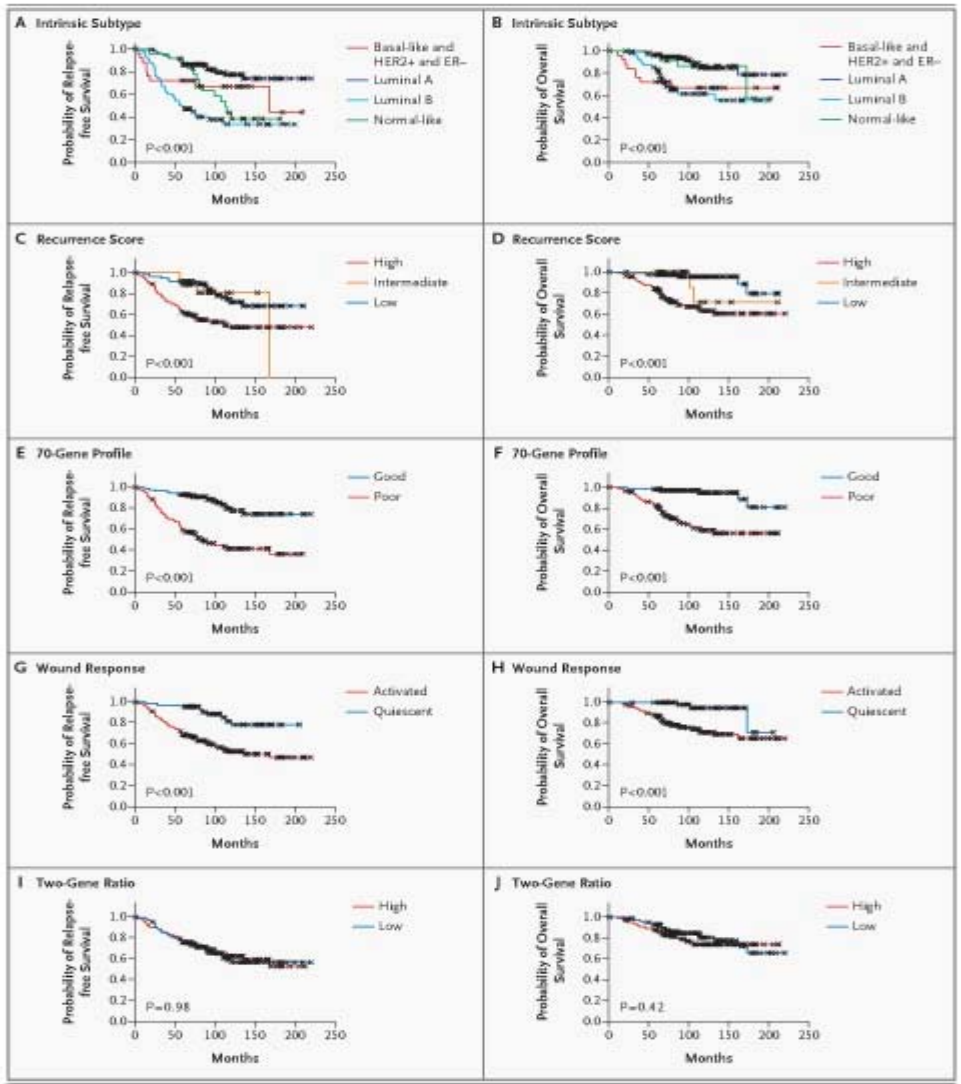
CONCLUSIONS

Even though different gene sets were used for prognostication in patients with breast cancer, four of the five tested showed significant agreement in the outcome predictions for individual patients and are probably tracking a common set of biologic phenotypes.

From the Departments of Genetics (C.F., D.S.O., C.M.P.), Statistics and Operations Research (A.B.N.), and Pathology and Laboratory Medicine (C.M.P.), University of North Carolina at Chapel Hill and Lineberger Comprehensive Cancer Center, Chapel Hill; and the Divisions of Diagnostic Oncology (L.W., B.W., L.J.V.) and Radiotherapy (D.S.A.N.), the Netherlands Cancer Institute, Amsterdam. Address reprint requests to Dr. Perou at Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Campus Box 7295, Chapel Hill, NC 27599, or at cperou@med.unc.edu.

Drs. Fan and Oh contributed equally to this article.

N Engl J Med 2006;355:560-9.



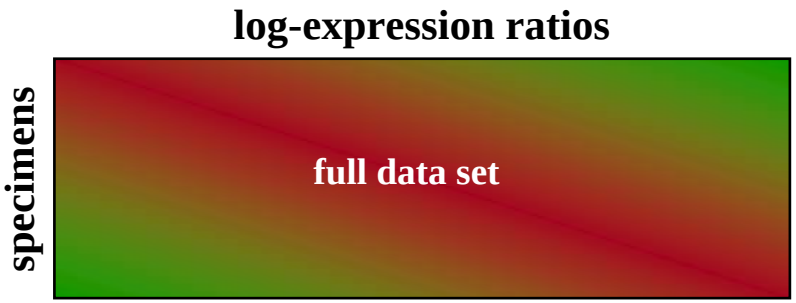
Evaluating a Classifier

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data
- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy

Split-Sample Evaluation

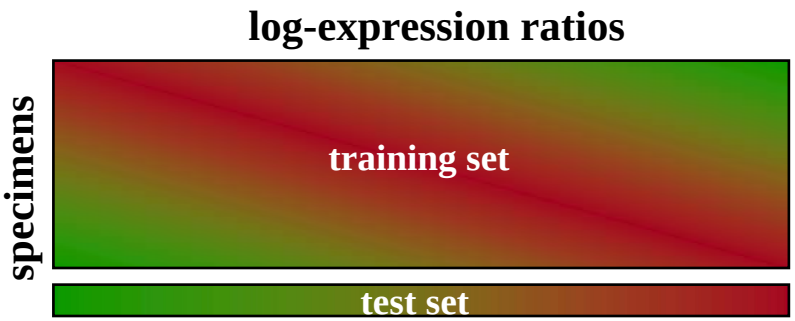
- Training-set
  - Used to select features, select model type, determine parameters and cut-off thresholds
- Test-set
  - Withheld until a *single* model is *fully* specified using the training-set.
  - Fully specified model is applied to the expression profiles in the test-set to predict class labels.
  - Number of errors is counted

Non-Cross-Validated Prediction



1. Prediction rule is built using full data set.
2. Rule is applied to each specimen for class prediction.

Cross-Validated Prediction (Leave-One-Out Method)



1. Full data set is divided into training and test sets (test set contains 1 specimen).
2. Prediction rule is built from scratch using the training set.
3. Rule is applied to the specimen in the test set for class prediction.
4. Process is repeated until each specimen has appeared once in the test set.

- Cross validation is only valid if the test set is not used in any way in the development of the model. Using the complete set of samples to select genes violates this assumption and invalidates cross-validation.
- With proper cross-validation, the model must be developed *from scratch* for each leave-one-out training set. This means that feature selection must be repeated for each leave-one-out training set.
  - Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data. Journal of the National Cancer Institute 95:14-18, 2003.

- The cross-validated estimate of misclassification error is an estimate of the prediction error for model fit using specified algorithm to full dataset

Myth

- Split sample validation is superior to LOOCV for estimating prediction error

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Prediction Error Estimation: A Comparison of Resampling Methods

Annette M. Molinaro [ab]*, Richard Simon [c], Ruth M. Pfeiffer [a]

[a]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, NCI, NIH, Rockville, MD 20852, [b]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, [c]Biometric Research Branch, Division of Cancer Treatment and Diagnostics, NCI, NIH, Rockville, MD 20852

ABSTRACT

Motivation: In genomic studies, thousands of features are collected on relatively few samples. One of the goals of these studies is to build classifiers to predict the outcome of future observations. There are three inherent steps to this process: feature selection, model selection, and prediction assessment. With a focus on prediction assessment, we compare several methods for estimating the 'true' prediction error of a prediction model in the presence of feature selection.

Results: For small studies where features are selected from thousands of candidates, the resubstitution and simple split-sample estimates are seriously biased. In these small samples, leave-one-out (LOOCV), 10-fold cross-validation (CV), and the .632+ bootstrap have the smallest bias for diagonal discriminant analysis, nearest neighbor, and classification trees. LOOCV and 10-fold CV have the smallest bias for linear discriminant analysis. Additionally, LOOCV, 5- and 10-fold CV, and the .632+ bootstrap have the lowest mean square error. The .632+ bootstrap is quite biased in small sample sizes with strong signal to noise ratios. Differences in performance among resampling methods are reduced as the number of specimens available increase.

Availability: A complete compilation of results in tables and figures is available in Molinaro *et al.* (2005). R code for simulations and analyses is available from the authors.

Contact: annette.molinaro@yale.edu

1 INTRODUCTION

In genomic experiments one frequently encounters high dimensional data and small sample sizes. Microarrays simultaneously monitor expression levels for several thousands of genes. Proteomic profiling studies using SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) measure size and charge of proteins and protein fragments by mass spectroscopy, and result in up to 15,000 intensity levels at prespecified mass values for each spectrum. Sample sizes in such experiments are typically less than 100.

In many studies observations are known to belong to predetermined classes and the task is to build predictors or classifiers for new observations whose class is unknown. Deciding which genes or proteomic measurements to include in the prediction is called *feature selection* and is a crucial step in developing a class predictor. Including too many noisy variables reduces accuracy of the prediction and may lead to over-fitting of data, resulting in promising but often non-reproducible results (Ransohoff, 2004).

Another difficulty is model selection with numerous classification models available. An important step in reporting results is assessing the chosen model's error rate, or generalizability. In the absence of independent validation data, a common approach to estimating predictive accuracy is based on some form of resampling the original data, e.g., cross-validation. These techniques divide the data into a learning set and a test set and range in complexity from the popular learning-test split to v-fold cross-validation, Monte-Carlo v-fold cross-validation, and bootstrap resampling. Few comparisons of standard resampling methods have been performed to date, and all of them exhibit limitations that make their conclusions inapplicable to most genomic settings. Early comparisons of resampling techniques in the literature are focussed on model selection as opposed to prediction error estimation (Breiman and Spector, 1992; Burman, 1989). In two recent assessments of resampling techniques for error estimation (Braga-Neto and Dougherty, 2004; Efron, 2004), feature selection was not included as part of the resampling procedures, causing the conclusions to be inappropriate for the high-dimensional setting.

We have performed an extensive comparison of resampling methods to estimate prediction error using simulated (large signal to noise ratio), microarray (intermediate signal to noise ratio) and proteomic data (low signal to noise ratio), encompassing increasing sample sizes with large numbers of features. The impact of feature selection on the performance of various cross validation methods is highlighted. The results elucidate the "best" resampling techniques for

*to whom correspondence should be addressed

BRB-ArrayTools

- Contains analysis tools that I have selected as valid and useful
- Analysis wizzard and multiple help screens for biomedical scientists
- Imports data from all platforms and major databases

Predictive Classifiers in BRB-ArrayTools

- Classifiers
  - Diagonal linear discriminant
  - Compound covariate
  - Bayesian compound covariate
  - Support vector machine with inner product kernel
  - K-nearest neighbor
  - Nearest centroid
  - Shrunken centroid (PAM)
  - Random forrest
  - Tree of binary classifiers for k-classes
- Survival risk-group
  - Supervised pc's
- Feature selection options
  - Univariate t/F statistic
  - Hierarchical variance option
  - Restricted by fold effect
  - Univariate classification power
  - Recursive feature elimination
  - Top-scoring pairs
- Validation methods
  - Split-sample
  - LOOCV
  - Repeated k-fold CV
  - .632+ bootstrap

BRB-ArrayTools

- Extensive built-in gene annotation and linkage to gene annotation websites
- Extensive gene-set enrichment tools for integrating gene expression with pathways and other biological information
- Publicly available for non-commercial use
  - http://linus.nci.nih.gov/brb

BRB-ArrayTools

December 2006

- 6635 Registered users
- 1938 Distinct institutions
- 68 Countries
- 311 Citations

Clinical Trial Design for Evaluating Medical Utility of a Predictive Biomarker for Use in Selecting Available Treatments

Studies Developing Gene Expression Profile Classifiers Should be Viewed as Analogous to Phase II Trials Requiring Phase III Validation

Limitations to Internal Validation

- Confounding by sample handling or assay effects
  - Cases collected and assayed at different times than controls
- Failure to incorporate important sources of variability
  - Assay variability
  - Tissue handling
  - Tumor heterogeneity
- Limited size of developmental study
- Problems of design and analysis in developmental study

Independent Data Validation

- From different clinical centers
- Specimens assayed at different time from training data
- Reproducibility of assay for individual tumors demonstrated to clinical reference laboratory standards
- Positive and negative samples collected in the same way
- Study sufficiently large to give precise estimates of sensitivity and specificity of the classifier
- The validation study is prospectively planned
  - patient selection pre-specified to address a therapeutically relevant question
  - endpoints and hypotheses pre-specified
  - predictor fully pre-specified
  - Study addresses assay reproducibility
  - Specimens may be either prospective or archived

Adequate External Validation Studies are Rarely Performed

- They are expensive and require multi-center cooperation
- They require demonstration of assay reproducibility
- The financial incentives for developing and validating PG classifiers of existing treatments are limited

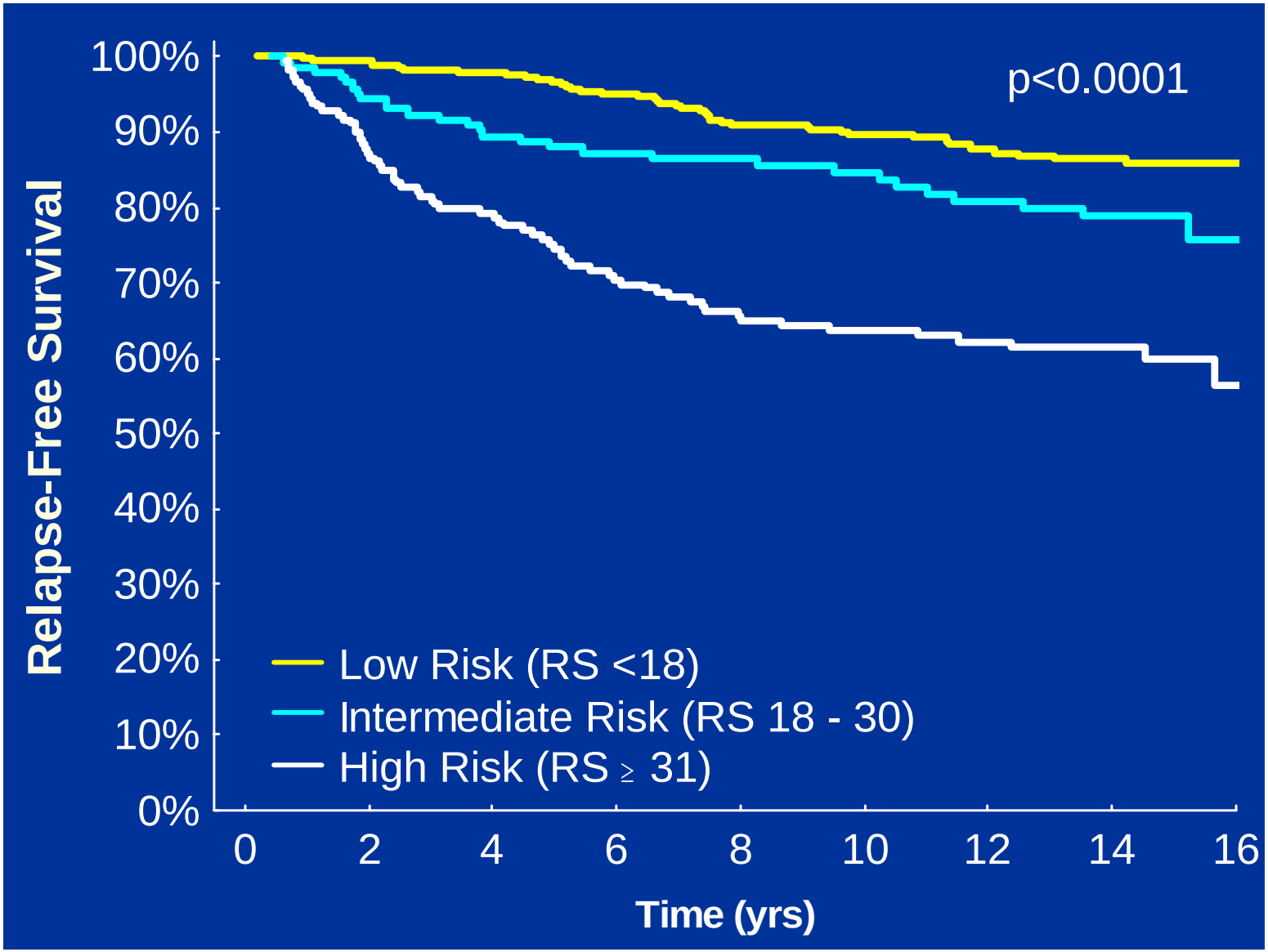
Validation Study
Node negative Breast Cancer

- Prospective study design
- Samples collected and archived from patients with node negative ER+ breast cancer receiving TAM
- Apply single, fully specified multi-gene predictor of outcome to samples and categorize each patient as good or poor prognosis
- Are long-term outcomes for patients in good prognosis group sufficiently good to withhold chemotherapy?

Prospectively Planned Validation Using Archived Materials
Oncotype-Dx

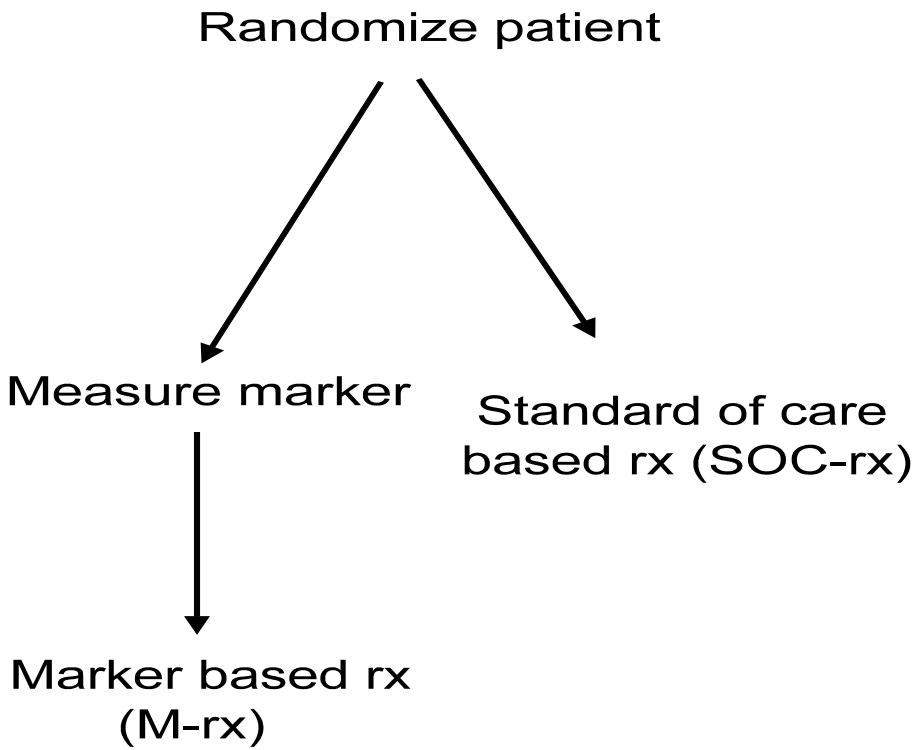
- Fully specified classifier applied prospectively to frozen specimens from NSABP B14 patients who received Tamoxifen for 5 years
- Long term follow-up available
- Good risk patients had very good relapse-free survival

B-14 Results—Relapse-Free Survival



Prospective Validation Designs

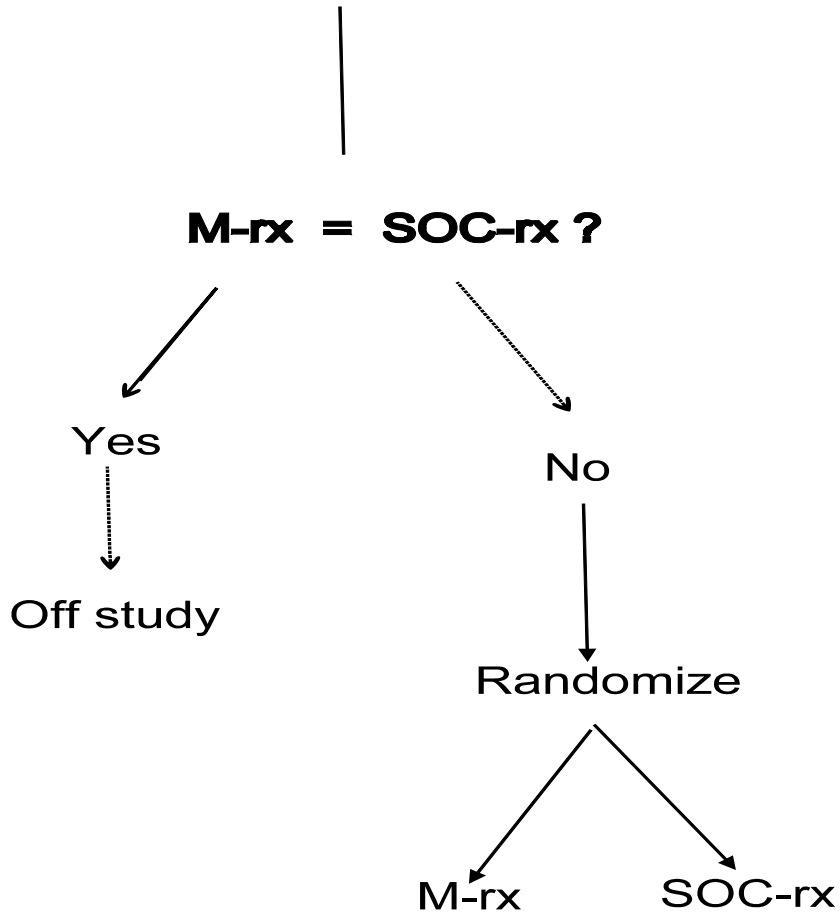
- Randomize patients to standard of care vs classifier determined therapy
- Gold standard but rarely performed
  - Very inefficient


Randomize patient
Measure marker
Standard of care based rx (SOC-rx)
Marker based rx (M-rx)

Proportion of Patients Marker +	Approximate Number of Events Required
20%	5200
33%	1878
50%	820

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

Determine marker based rx (M-rx) and standard of care based rx (SOC-rx)



Prospective Intergroup Study

- OncotypeDx risk score <15
  - Tam alone
- OncotypeDx risk score >30
  - Tam + Chemo
- OncotypeDx risk score 15-30
  - Randomize to Tam vs Tam + Chemo

Collaborators

- Alain Dupuy
- Boris Freidlin
- Wenyu Jiang
- Aboubakar Maitournam
- Annette Molinaro
- Michael Radmacher
- Joanna Shih
- Sue Jane Wang
- Yingdong Zhao
- BRB-ArrayTools Development Team

Using Genomic Classifiers In Clinical Trials

•Dupue A and Simon R. Critical review of published microarray studies for clinical outcome and guidelines for statistical analysis and reporting, Journal of the National Cancer Institute 99:147-57, 2007
.

•Dobbin K and Simon R. Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics 8:101-117, 2007.

•Simon R. Development and validation of therapeutically relevant predictive classifiers using gene expression profiling, Journal of the National Cancer Institute 98:1169-71, 2006.

•Simon R. Validation of pharmacogenomic biomarker classifiers for treatment selection. Cancer Biomarkers 2:89-96, 2006.

•Simon R. Guidelines for the design of clinical studies for development and validation of therapeutically relevant biomarkers and biomarker classification systems. In Biomarkers in Breast Cancer, Hayes DF and Gasparini G, pp 3-15, Humana Press, 2006.

•Simon R. A checklist for evaluating reports of expression profiling for treatment selection. Clinical Advances in Hematology and Oncology 4:219-224, 2006.

•Simon R. Identification of pharmacogenomic biomarker classifiers in drug development. In Pharmacogenomics, Anti-cancer Drug Discovery and Response, F Innocenti (ed), Humana Press (In Press).

•Simon R, Lam A, Li MC, et al. Analysis of gene expression data using BRB-ArrayTools, Cancer Informatics 2:1-7, 2006

•Simon R. New challenges for 21st century clinical trials, Controlled Clinical Trials (In Press).

•Simon R. Development and validation of biomarker classifiers for treatment selection, Journal of Statistical Planning and Inference (In Press).

Using Genomic Classifiers In Clinical Trials

.

•Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 10:6759-63, 2004.

•Maitnourim A and Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine 24:329-339, 2005.

•Simon R. When is a genomic classifier ready for prime time? Nature Clinical Practice – Oncology 1:4-5, 2004.

•Simon R. An agenda for Clinical Trials: clinical trials in the genomic era. Clinical Trials 1:468-470, 2004.

•Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. Journal of the National Cancer Institute 97:866-867, 2005..

•Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology 23:7332-41,2005.

•Freidlin B and Simon R. Adaptive signature design. Clinical Cancer Research 11:7872-78, 2005.

•Simon R. and Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases, The Pharmacogenomics Journal 6:166-73, 2006.

•Trepicchio WL, Essayan D, Hall ST, Schechter G, Tezak Z, Wang SJ, Weinreich D, Simon R. Designing prospective clinical pharmacogenomic trials- Effective use of genomic biomarkers for use in clinical decision-making. The Pharmacogenomics Journal 6:89-94,2006.