Cancer Clinical Trials in the Genomic Era

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

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

- Powerpoint presentation
- Reprints & Technical Reports
- BRB-ArrayTools software
- Interactive sample size planning for targeted clinical trials

Genomics Can Influence

- New treatments developed
- Phase I/II development
- Target patient population

“Biomarkers”

- Surrogate endpoints
  - A measurement made on a patient before, during and after treatment to determine whether the treatment is working
- Predictive classifier
  - A measurement made before treatment to predict whether a particular treatment is likely to be beneficial

Surrogate Endpoints

- It is extremely difficult to properly validate a biomarker as a surrogate for clinical benefit. It requires a series of randomized trials with both the candidate biomarker and clinical outcome measured

- Biomarkers can be useful in phase I/II studies and need not be validated as surrogates for clinical benefit
- Unvalidated surrogates can also be used for early termination of phase III trials. The trial should continue accrual and follow-up to evaluate true endpoint if treatment effect on partial surrogate is sufficient.

Predictive Classifiers

- Most cancer treatments benefit only a minority of patients to whom they are administered
  - Particularly true for molecularly targeted drugs
- Being able to predict which patients are likely to benefit would
  - save patients from unnecessary toxicity
  - enhance their chance of receiving a drug that helps them
  - Reduce the size of phase III clinical trials
  - Help control medical costs

Oncology Needs Predictive Markers not Prognostic Factors

- Many prognostic factor studies use a convenience sample of patients for whom tissue is available. Generally the patients are too heterogeneous to support therapeutically relevant conclusions

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

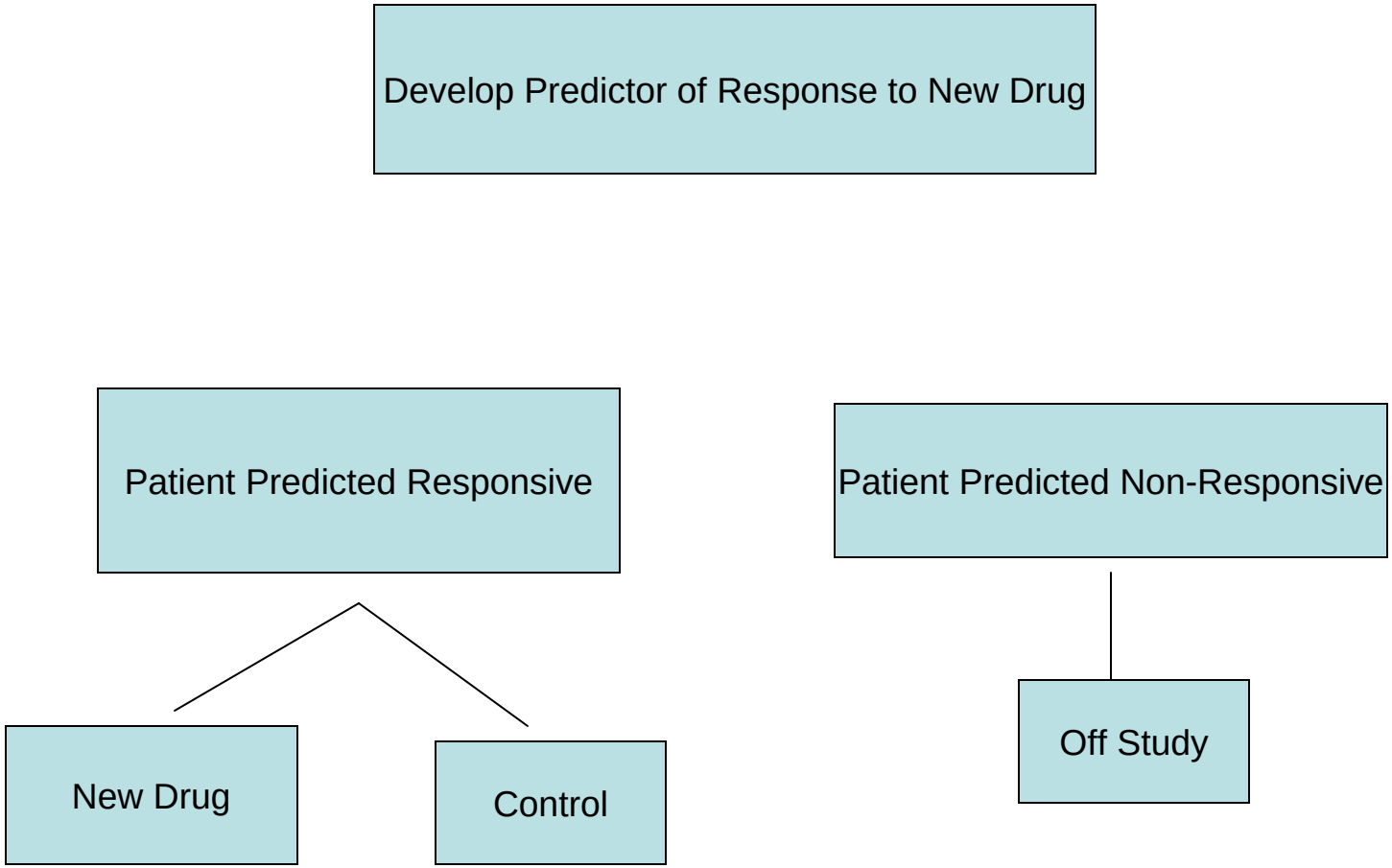
- 939 articles on “prognostic markers” or “prognostic factors” in breast cancer in past 20 years
- ASCO guidelines only recommend routine testing for ER, PR and HER-2 in breast cancer
- “With the exception of ER or progesterone receptor expression and HER-2 gene amplification, there are no clinically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

- Targeted clinical trials can be much more efficient than untargeted clinical trials, if we know who to target

- In new drug development, the role of a predictive classifier is to select a target population for treatment
  - The focus should be on evaluating the new drug in a population defined by a predictive classifier, not on “validating” the classifier

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

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

Randomized Ratio
(normal approximation)

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

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

- δ_1= rx effect in marker + patients
- δ_0= rx effect in marker - patients
- λ=proportion of marker + patients
- If δ_0=0, RandRat = $1/\lambda^2$
- If $\delta_0 = \delta_1/2$, RandRat = $4/(\lambda+1)^2$

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

Proportion Marker Positive	No Treatment Benefit for Marker Negative Patients	Treatment Benefit for Marker Negative Patients is Half That for Marker Positive Patients
0.75	1.78	1.31
0.5	4	1.78
0.25	16	2.56

Screened Ratio

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

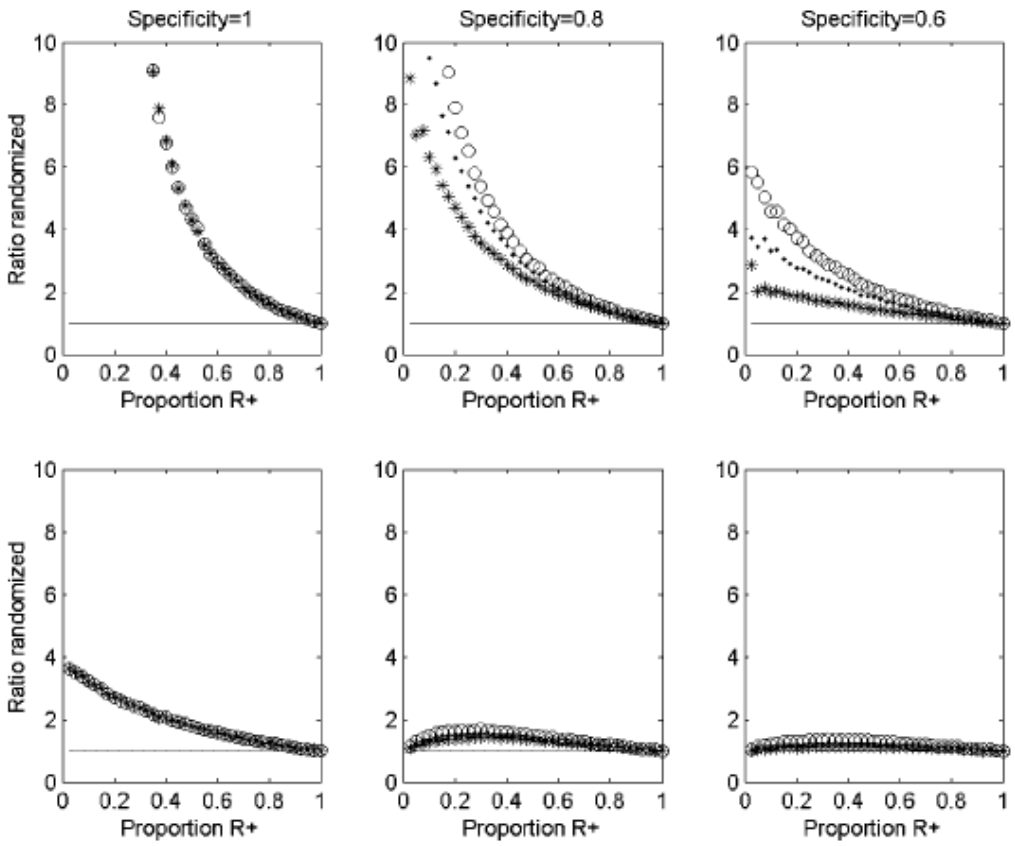


Figure 1. Ratio of number randomized for untargeted versus targeted designs. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ○ Sensitivity = 1; • Sensitivity = 0.8; ⋆ Sensitivity = 0.6.

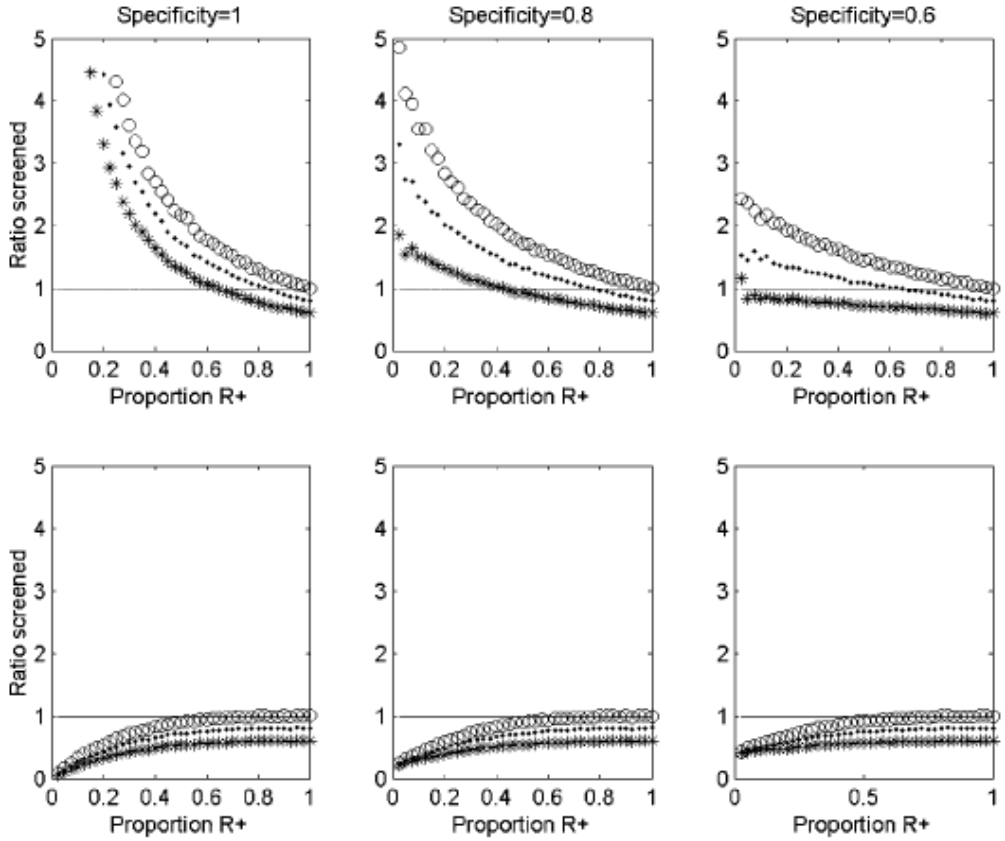


Figure 2. Ratio of number randomized for untargeted design to number screened for targeted design. Upper panel: no treatment effect for R− patients. Lower panel: treatment effect for R− patients half that of R+ patients. ○ Sensitivity = 1; • Sensitivity = 0.8; ⋆ Sensitivity = 0.6.

- For Trastuzumab, even a relatively poor assay enabled conduct of a targeted phase III trial which was crucial for establishing effectiveness
- Recent results with Trastuzumab in early stage breast cancer show dramatic benefits for patients selected to express Her-2

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

Interactive Software for Evaluating a Targeted Design

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

Biometric Research Branch home page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://linus.nci.nih.gov/~brb/ Go Links

Google west hawaii cancer symposium Search Options west hawaii cancer symposium

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

Internet

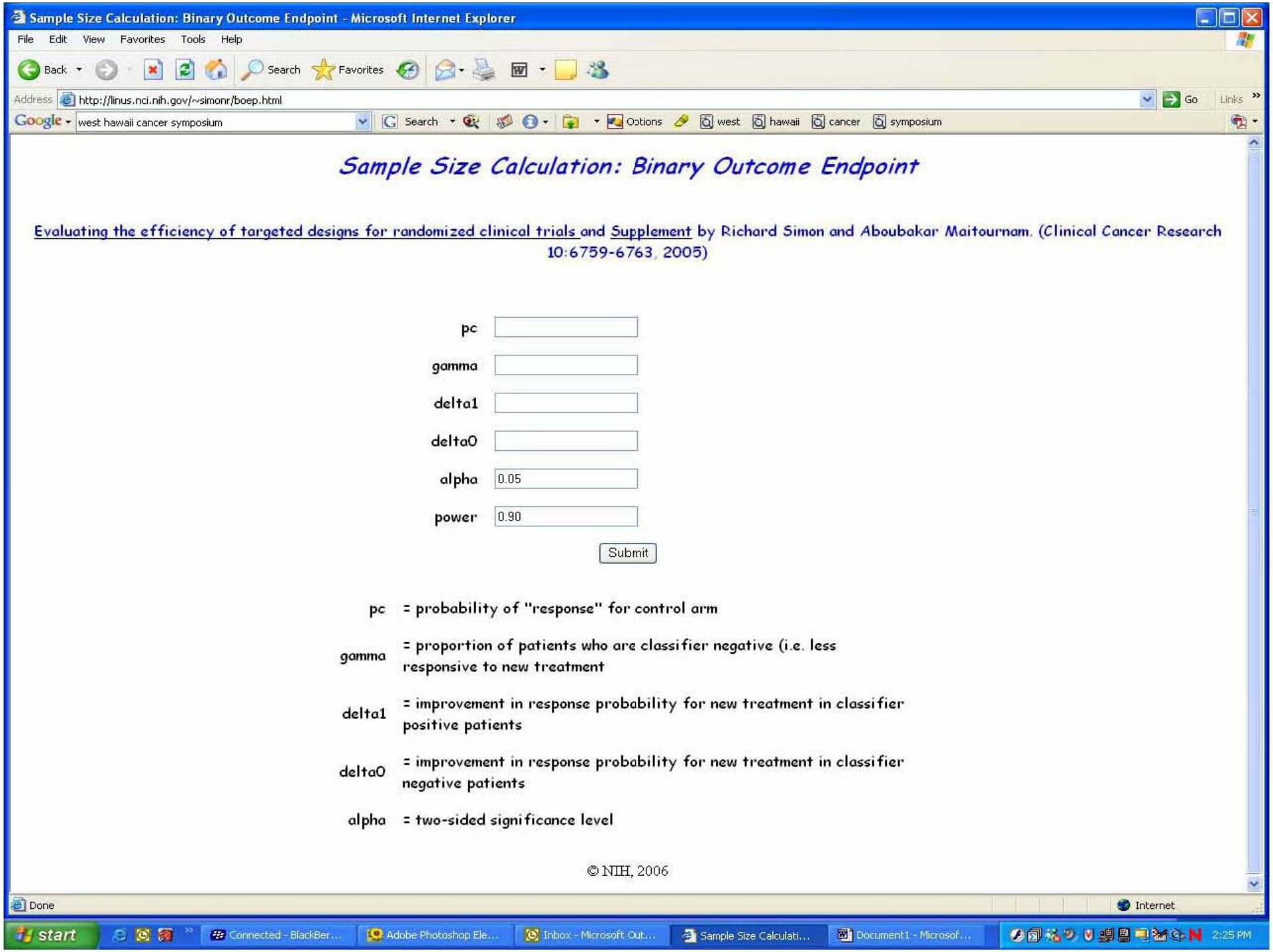
start Connected - BlackBer... Adobe Photoshop Ele... Inbox - Microsoft Out... Biometric Research B... Document1 - Microsof... 2:25 PM

Sample Size Calculation for Randomized Clinical Trials

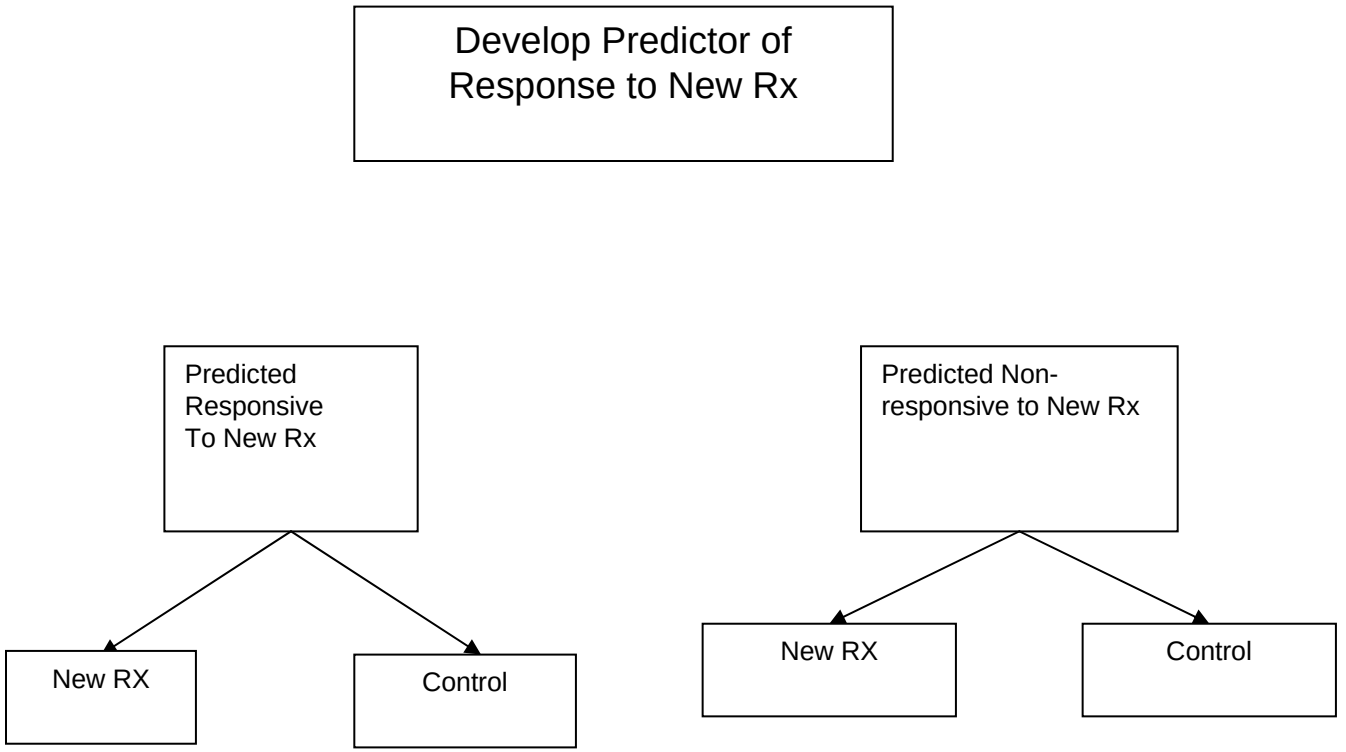
- **Optimal Two-Stage Phase II Design**

- **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 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://linus.nci.nih.gov/~simonr/boep.html
Go Links
Google west hawaii cancer symposium
Search Options west hawaii cancer symposium
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
© NIH, 2006
Done
Internet
start
Connected - BlackBer...
Adobe Photoshop Ele...
Inbox - Microsoft Out...
Sample Size Calculati...
Document1 - Microsof...
2:25 PM

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.

- The purpose of the RCT is to evaluate treatment T vs C overall and for the pre-defined subset; not to re-evaluate the components of the classifier, or to modify or refine the classifier

Sample Size Planning for Design II

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

Developmental Strategy (IIb)

- 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 treatment effect at 0.05 level

Sample size Planning for IIb

- Accrue classifier + and - patients in a manner that enriches for classifier + patients until there are sufficient classifier + patients for standard power at significance level 0.05 for detecting large treatment effect D
- If treatment is found effective in classifier + patients, continue accrual of - patients for standard power at significance level 0.05 for detecting usual size treatment effect d representing minimal useful clinical utility

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
  - 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

Development of Genomic Classifiers

- Single gene or protein based on knowledge of therapeutic target
- Single gene or protein culled from set of candidate genes identified based on imperfect knowledge of therapeutic target
- Empirically determined based on correlating gene expression 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 Empirical Gene Expression Based Classifier

- 20-30 phase II responders are needed to compare to non-responders in order to develop signature for predicting response
  - Dobbin KK, Simon RM. Sample size planning for developing classifiers using high dimensional DNA microarray data, Biostatistics (In Press); available at http://linus.nci.nih.gov

Development of Empirical Gene Expression Based Classifier

- A signature of response to the new drug may not represent a signature of preferential benefit from a regimen containing the new drug versus a control regimen

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
    - Genes selected based on interaction between expression level and treatment effect (E vs C)
    - Weighted voting classifier used
  - 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

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

Myths about the Development of Predictive Classifiers using Gene Expression Profiles

Myth

- Microarray studies are exploratory with no hypotheses or objectives

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

Myth

- Cluster analysis is a useful for analysis of most microarray studies

Class Comparison and Class Prediction

- Not clustering problems
- Supervised methods should be used

Myth

- Development of good predictive classifiers is not possible with >1000 genes and <100 cases
- Predictive models should be reproducible on independent data

- Much of the conventional wisdom of statistical analysis is focused on inference, not on prediction
- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy
- Predictive models should predict accurately for independent data; the model itself need not be reproducibly derivable on independent data
- Most statistical methods were not developed for prediction problems and particularly not for prediction problems with >10,000 variables and <100 cases
- Accurate prediction is possible for p>>n problems if there are sufficient informative genes and new approaches to model development are used

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.

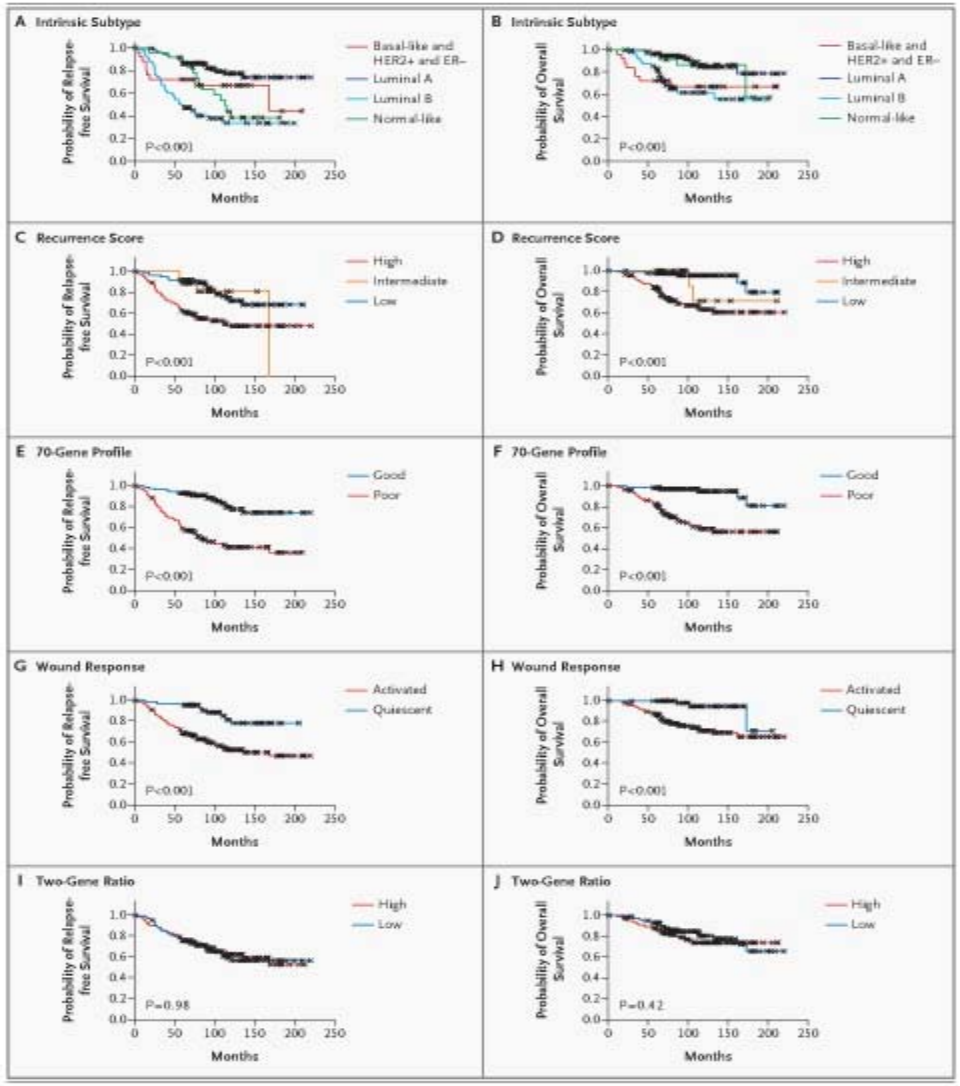
CONCLUSIONS

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.



Myth

- 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 generally indicate that simpler methods work as well or better for microarray problems because they avoid overfitting the data.

A set of genes is not a classifier

- Gene selection
- Mathematical function for mapping from multivariate gene expression domain to prognostic or diagnostic classes
- Weights and other parameters including cut-off thresholds for risk scores

Simple and Effective Classifiers

- Select genes that are individually correlated with outcome
- Linear classifiers
  - Diagonal LDA, Compound covariate predictor, Weighted voting classifier, Linear Support vector machines
- Nearest neighbor and shrunken centroid classifiers

Feature Selection

- 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
    - Methods for class prediction are different than those for class comparison
  - The accuracy of the significance test used for feature selection is not of major importance as identifying differentially expressed genes is not the ultimate objective

Feature Selection

- Small subset of genes which together give most accurate predictions
  - Combinatorial optimization algorithms
    - Genetic algorithms
- Little evidence that complex feature selection is useful in microarray problems
  - Failure to compare to simpler methods
  - Some published complex methods for selecting combinations of features do not appear to have been properly evaluated

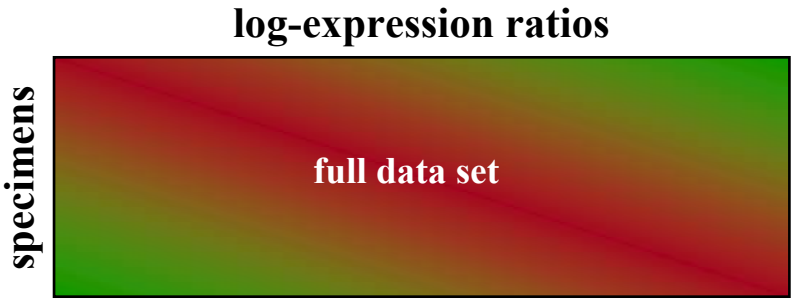
Evaluating a Classifier

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

Split-Sample Evaluation

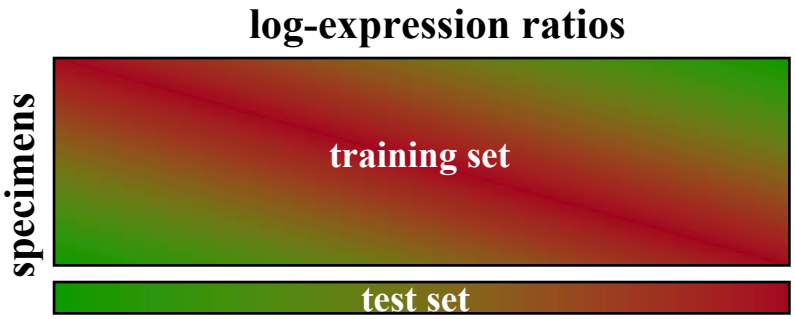
- Training-set
  - Used to select features, select model type, determine parameters and cut-off thresholds
- Test-set
  - Withheld until a *single* model is *fully* specified using the training-set.
  - Fully specified model is applied to the expression profiles in the test-set to predict class labels.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

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 or 10-fold CV 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

Limitations to Internal Validation

- Sample handling and assay conduct are performed under controlled conditions that do not incorporate real world sources of variability
- Developmental studies are generally small
- Predictive accuracy is often not clinical utility

External Validation

- From different clinical centers
- Specimens assayed at different time from training data
- Samples handled and assayed blinded from clinical outcome
- Study sufficiently large to give precise estimates of sensitivity and specificity of the classifier
- Study addresses clinical utility of using the genomic classifier compared to using standard practice guidelines

Myth

- Huge sample sizes are needed to develop effective predictive classifiers

Sample Size Planning References

- K Dobbin, R Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6:27-38, 2005
- K Dobbin, R Simon. Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics (In Press)

Sample Size Planning for Classifier Development

- The expected value (over training sets) of the probability of correct classification PCC(n) should be within γ of the maximum achievable PCC(◎)

Probability Model

- Two classes
- Log expression or log ratio MVN in each class with common covariance matrix
- m differentially expressed genes
- p-m noise genes
- Expression of differentially expressed genes are independent of expression for noise genes
- All differentially expressed genes have same inter-class mean difference 2δ
- Common variance for differentially expressed genes and for noise genes

Classifier

- Feature selection based on univariate t-tests for differential expression at significance level α
- Simple linear classifier with equal weights (except for sign) for all selected genes. Power for selecting each of the informative genes that are differentially expressed by mean difference 2δ is $1-\beta(n)$

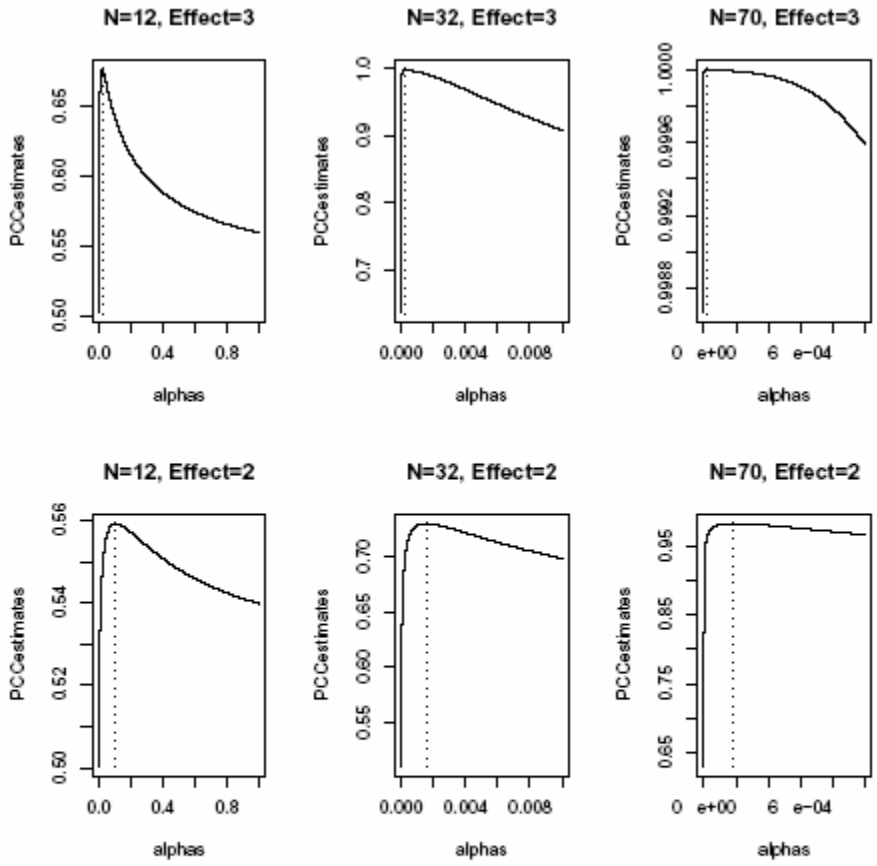


Figure 1: Plots of the estimated PCC as a function of α, plotted for various values of n, based on Equation 10. In each plot, "Effect" is defined as $\frac{2\delta}{\sigma}$, $m = 10$ is the number of differentially

Optimal significance level cutoffs for gene selection. 50 differentially expressed genes out of 22,000 genes on the microarrays

$2\delta/\sigma$	**n=10**	**n=30**	**n=50**
1	0.167	0.003	0.00068
1.25	0.085	0.0011	0.00035
1.5	0.045	0.00063	0.00016
1.75	0.026	0.00036	0.00006
2	0.015	0.0002	0.00002

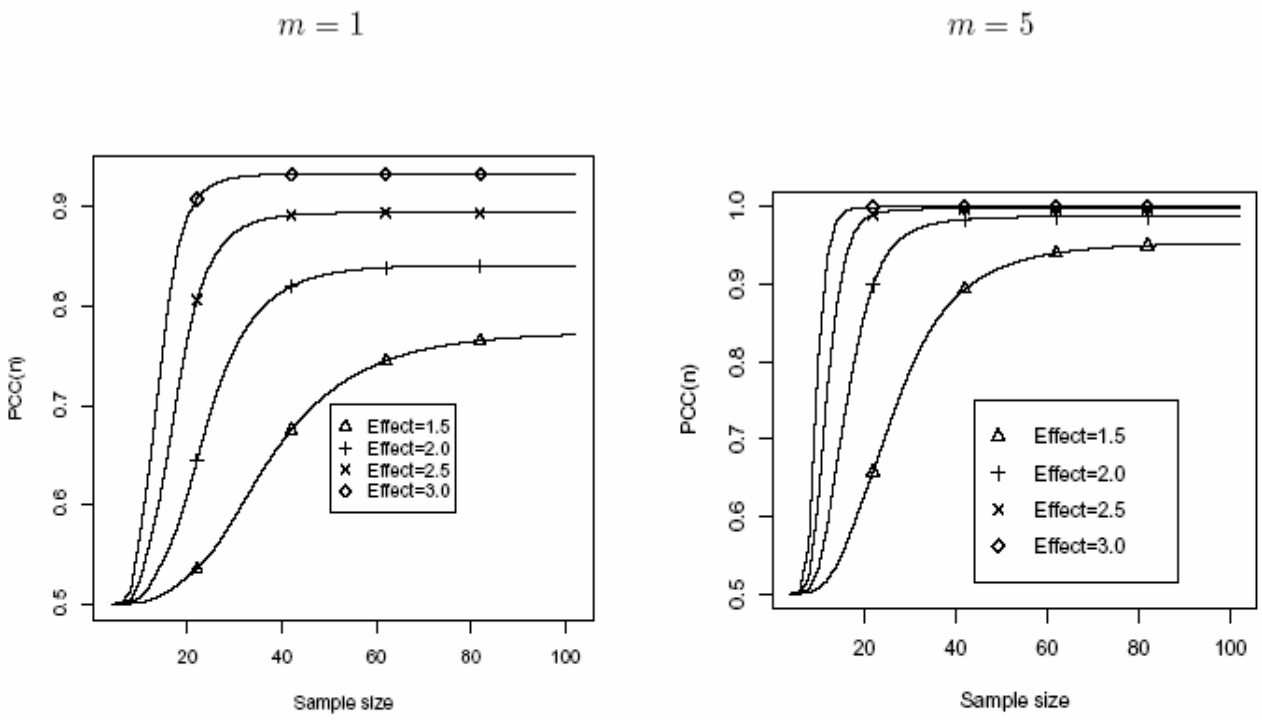
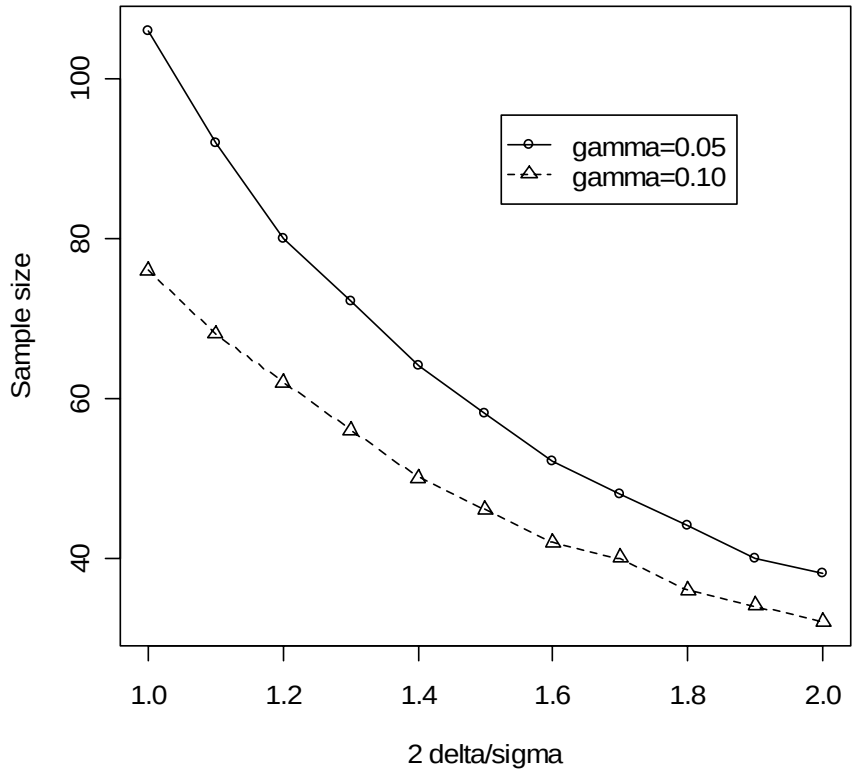


Figure 3: Left plot is $m = 1$ and right plot is $m = 5$. $p = 10,000$. Plot of sample size versus probability of correct classification for various values of the effect size $2\delta/\sigma$. Gene independence is assumed. $PCC(n)$ use optimal α. Population assumed evenly split between the classes, so $p_1 = 1/2$.

Sample size as a function of effect size (log-base 2 fold-change between classes divided by standard deviation). Two different tolerances shown, . Each class is equally represented in the population. 22000 genes on an array.



Class Comparison
2 equal size classes

$$n = 4\sigma^2(z_{\alpha/2} + z_{\beta})^2/\delta^2$$

where δ = mean log-ratio difference between classes

σ = within class standard deviation of biological replicates

$z_{\alpha/2}$, z_{β} = standard normal percentiles

- Choose α small, e.g. $\alpha = .001$
- Use percentiles of t distribution for improved accuracy

- π = proportion of genes on array that are differentially expressed between classes
- N = number of genes on the array
- FD = expected number of false discoveries
- TD = expected number of true discoveries
- FDR = FD/(FD+TD)

- FD = $\alpha(1-\pi)N$
- TD = $(1-\beta)\ \pi N$
- FDR = $\alpha(1-\pi)N/\{\alpha(1-\pi)N + (1-\beta)\ \pi N\}$
- $= 1/\{1 + (1-\beta)\pi/\alpha(1-\pi)\}$

Controlling Expected False Discovery Rate

π	α	β	FDR
0.01	0.001	0.10	9.9%
	0.005		35.5%
0.05	0.001		2.1%
	0.005		9.5%

Total Number of Samples for Two Class Comparison

α	β	δ	σ	Samples Per Class
0.001	0.05	1 (2-fold)	0.5 human tissue	13
			0.25 transgenic mice	6 (t approximation)

Number of Events Needed to Detect Gene Specific Effects on Survival

- σ = standard deviation in log2 ratios for each gene
- Ω = hazard ratio (>1) corresponding to 2-fold change in gene expression

$$\left[\frac{z_{1-\alpha/2} + z_{1-\beta}}{\sigma \log_2 \delta} \right]^2$$

Number of Events Required to Detect Gene Specific Effects on Survival ♋=0.001, ♌=0.05

Hazard Ratio ♎	♦	Events Required
2	**0.5**	**26**
1.5	**0.5**	**76**

Selected Features of BRB-ArrayTools linus.nci.nih.gov/brb

- Gene finding
  - Multivariate permutation tests
  - Fast SAM
  - t/F tests with hierarchical variance model
  - Class comparison, survival comparison, quantitative trait correlation
- Extensive gene annotation
- Gene set comparison analysis
  - GO, pathways, signatures, TF targets, protein domains
- Analysis of variance
  - Fixed, mixed, time-course, complex 2-color designs

Selected Features of BRB-ArrayTools

- Class prediction
  - DLDA, CCP, Nearest Neighbor, Nearest Centroid, Shrunken Centroids, SVM, Random Forests,Top scoring pairs, naïve Bayesian classification
  - Complete LOOCV, k-fold CV, repeated k-fold, .632+ bootstrap
  - permutation significance of cross-validated error rate
- Survival risk group prediction
- R plug-ins

Conclusions

- New technology and biological knowledge make it increasingly feasible to identify which patients are most likely to benefit from a specified treatment
- “Predictive medicine” is feasible but does not mean “personalized treatment”
- Targeting treatment can greatly improve the therapeutic ratio of benefit to adverse effects
  - Smaller clinical trials needed
  - Treated patients benefit
  - Economic benefit for society

Conclusions

- Achieving the potential of new technology requires paradigm changes in focus and methods of “correlative science.”
- Achieving the potential of new technology requires paradigm changes in partnerships among industry, academia, NIH and FDA.
- Effective interdisciplinary research requires increased emphasis on cross education of laboratory, clinical and statistical scientists

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 previous developmental studies

Collaborators

- Kevin Dobbin
- Boris Freidlin
- Aboubakar Maitournam
- Annette Molinaro
- Ruth Pfeifer
- Michael Radmacher
- Yingdong Zhao

Simon R, Korn E, McShane L, Radmacher M, Wright G, Zhao Y. *Design and analysis of DNA microarray investigations*, Springer-Verlag, 2003.

Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511, 2002.

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.

Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery, Bioinformatics 18:1462-69, 2002; 19:803-810, 2003; 21:2430-37, 2005; 21:2803-4, 2005.

Dobbin K and Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6:27-38, 2005.

Dobbin K, Shih J, Simon R. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. Journal of the National Cancer Institute 95:1362-69, 2003.

Wright G, Simon R. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 19:2448-55, 2003.

Korn EL, Troendle JF, McShane LM, Simon R.Controlling the number of false discoveries. Journal of Statistical Planning and Inference 124:379-08, 2004.

Molinaro A, Simon R, Pfeiffer R. Prediction error estimation: A comparison of resampling methods. Bioinformatics 21:3301-7,2005.

Simon R. Using DNA microarrays for diagnostic and prognostic prediction. Expert Review of Molecular Diagnostics, 3(5) 587-595, 2003.

Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. British Journal of Cancer 89:1599-1604, 2003.

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(29), 2005.

Freidlin B and Simon R. Adaptive signature design. Clinical Cancer Research 11:7872-8, 2005.

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, Humana Press, pp 3-15, 2005.

Simon R and Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. The Pharmacogenomics Journal, 2006.