Using Predictive Biomarkers in the Design of Phase III Clinical Trials

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

Disclosure Information

AACR 2007 National Meeting

Dr. Richard Simon

I have no financial relationships to disclose.

I will not discuss off label use and/or investigational use in my presentation.

Biometric Research Branch 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
  - Treated patients benefit
  - Treatment more cost-effective for society

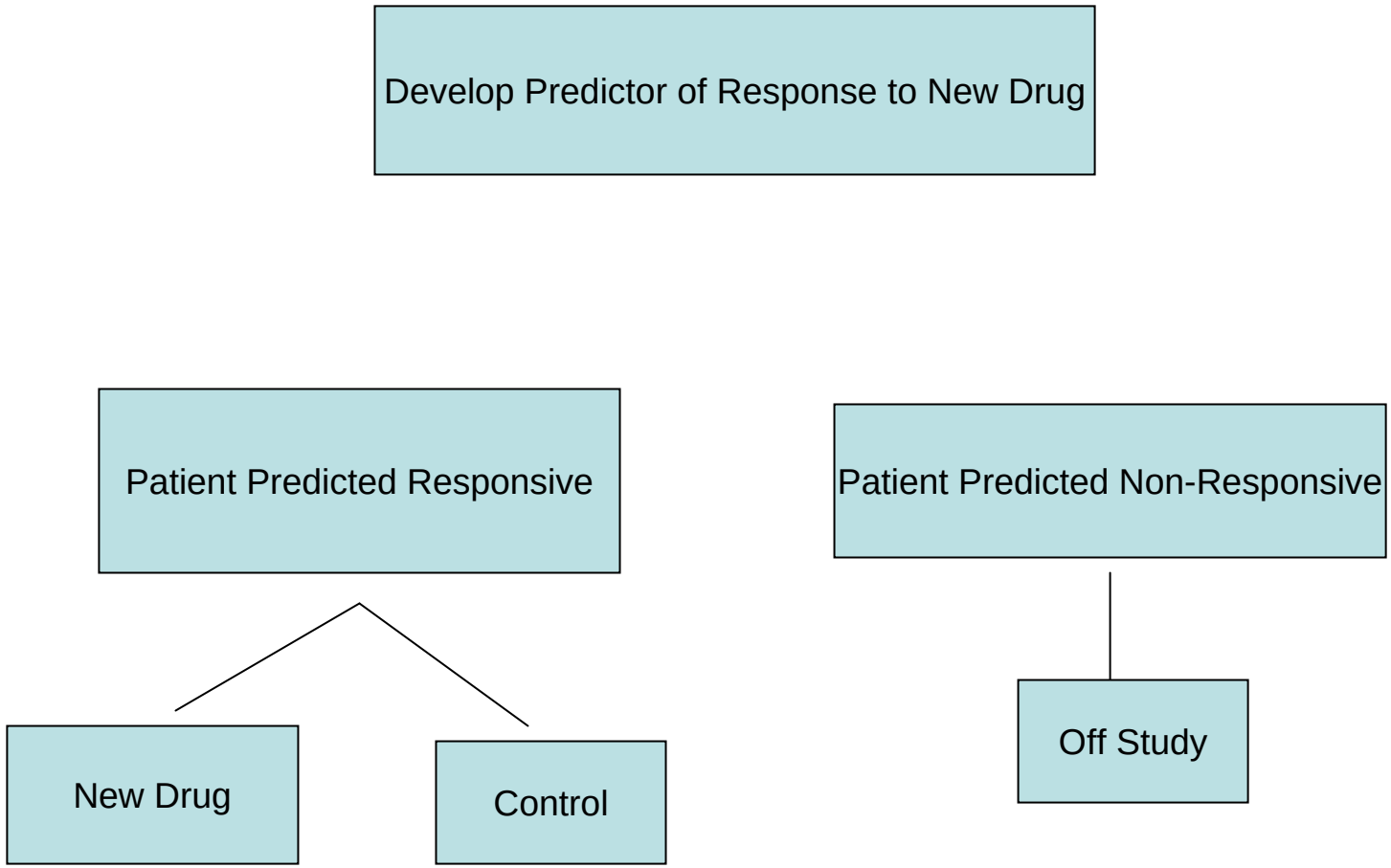
Medicine Needs Predictive Markers not Prognostic Factors

- Most prognostic factors are not used because they are not therapeutically relevant
- Most prognostic factor studies are not focused on a clear objective
  - they use a convenience sample of patients for whom tissue is available
  - often the patients are too heterogeneous to support therapeutically relevant conclusions

- 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 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; Correction 12:3229,2006
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine 24:329-339, 2005.

Two Clinical Trial Designs Compared

- 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

- δ_1 = treatment effect for Target + patients
- δ_0 = treatment effect for Target - patients

- Sensitivity = Prob{Assay+ | Target +}
- Specificity = Prob{Assay- | Target -}

Randomized Ratio

randomized: standard design / targeted design

sensitivity=specificity=0.9

Proportion Expressing 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

screened standard design / targeted design

sensitivity=specificity=0.9

Proportion Expressing Target	$\delta_0=0$	$\delta_0=\delta_1/2$
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

- 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

- 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

Comparison of Targeted to Untargeted Design Disease-Free Survival Endpoint

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

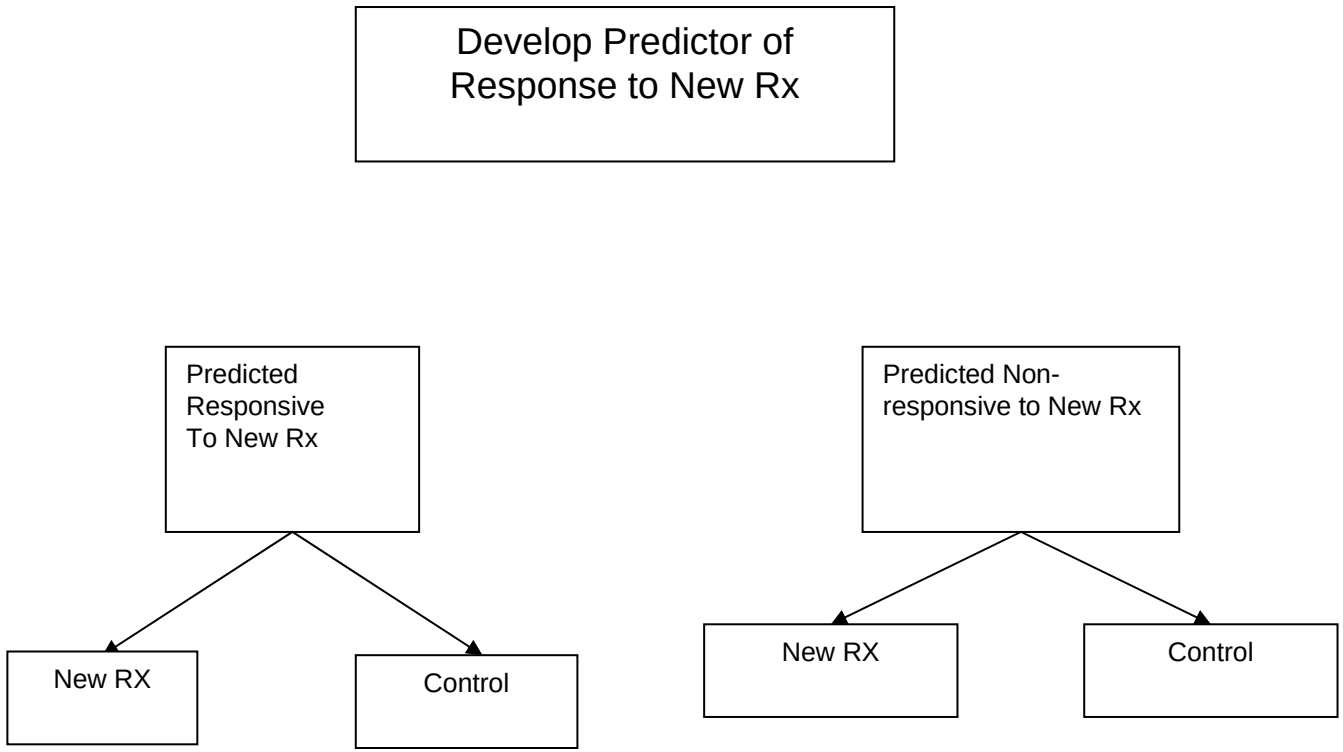
alpha = two-sided significance level

*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”

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

Sample Size Planning for Designs II and III

- II - Size for standard power (e.g. 0.9) for detecting usual treatment effect overall at significance level 0.04
- III - Size for standard power (e.g. 0.9) for detecting larger treatment effect in positive subset

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
  - 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 genes
  - 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.

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

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
- Survival or DFS endpoint
- Have identified a biomarker index
  - No threshold pre-determined
- Eligibility not restricted by biomarker index
- Is E superior to C overall or for patient subset defined by range of index?

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).

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 extensive analysis tools that I have selected as valid and useful
- Analysis wizard and multiple help screens for biomedical scientists
- Imports data from all platforms and major databases
- Extensive built-in gene annotation and linkage to gene annotation websites
- Publicly available for non-commercial use
  - http://linus.nci.nih.gov/brb

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

December 2006

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

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

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

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

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; Correction 12:3229, 2006.

•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.