

# Statistical Challenges for Predictive Oncology

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

# Biometric Research Branch Website

[brb.nci.nih.gov](http://brb.nci.nih.gov)

- Powerpoint presentations
- Reprints
- BRB-ArrayTools software
- Web based sample size planning for therapeutics and predictive biomarkers

# Prognostic & Predictive Biomarkers

- Predictive biomarkers
  - Measured before treatment to identify who is likely or unlikely to benefit from a particular treatment
- Prognostic biomarkers
  - Measured before treatment to indicate long-term outcome for patients untreated or receiving standard treatment

# Prognostic & Predictive Biomarkers

- Most cancer treatments benefit only a minority of patients to whom they are administered
- Being able to predict which patients are or are not likely to benefit would
  - Save patients from unnecessary toxicity, and enhance their chance of receiving a drug that helps them
  - Control medical costs
  - Improve the success rate of clinical drug development

# Prognostic & Predictive Biomarkers

- Single gene or protein measurement
  - ER protein expression
  - HER2 amplification
  - EGFR mutation
  - KRAS mutation
- Index or classifier that summarizes expression levels of multiple genes
  - OncotypeDx recurrence score

# Clinical Utility

- Biomarker benefits patients by improving treatment decisions
  - Identify patients who have very good prognosis on standard treatment and do not require more intensive regimens
  - Identify patients who are likely or unlikely to benefit from a specific regimen

## K-ras Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer

Christos S. Karapetis, M.D., Shirin Khambata-Ford, Ph.D., Derek J. Jonker, M.D., Chris J. O'Callaghan, Ph.D., Dongsheng Tu, Ph.D., Niall C. Tebbutt, Ph.D., R. John Simes, M.D., Haji Chalchal, M.D., Jeremy D. Shapiro, M.D., Sonia Robitaille, M.Sc., Timothy J. Price, M.D., Lois Shepherd, M.D.C.M., Heather-Jane Au, M.D., Christiane Langer, M.D., Malcolm J. Moore, M.D., and John R. Zalcberg, M.D., Ph.D.\*

### ABSTRACT

#### BACKGROUND

Treatment with cetuximab, a monoclonal antibody directed against the epidermal growth factor receptor, improves overall and progression-free survival and preserves the quality of life in patients with colorectal cancer that has not responded to chemotherapy. The mutation status of the *K-ras* gene in the tumor may affect the response to cetuximab and have treatment-independent prognostic value.

#### METHODS

We analyzed tumor samples, obtained from 394 of 572 patients (68.9%) with colorectal cancer who were randomly assigned to receive cetuximab plus best supportive care or best supportive care alone, to look for activating mutations in exon 2 of the *K-ras* gene. We assessed whether the mutation status of the *K-ras* gene was associated with survival in the cetuximab and supportive-care groups.

#### RESULTS

Of the tumors evaluated for *K-ras* mutations, 42.3% had at least one mutation in exon 2 of the gene. The effectiveness of cetuximab was significantly associated with *K-ras* mutation status ( $P=0.01$  and  $P<0.001$  for the interaction of *K-ras* mutation status with overall survival and progression-free survival, respectively). In patients with wild-type *K-ras* tumors, treatment with cetuximab as compared with supportive care alone significantly improved overall survival (median, 9.5 vs. 4.8 months; hazard ratio for death, 0.55; 95% confidence interval [CI], 0.41 to 0.74;  $P<0.001$ ) and progression-free survival (median, 3.7 months vs. 1.9 months; hazard ratio for progression or death, 0.40; 95% CI, 0.30 to 0.54;  $P<0.001$ ). Among patients with mutated *K-ras* tumors, there was no significant difference between those who were treated with cetuximab and those who received supportive care alone with respect to overall survival (hazard ratio, 0.98;  $P=0.89$ ) or progression-free survival (hazard ratio, 0.99;  $P=0.96$ ). In the group of patients receiving best supportive care alone, the mutation status of the *K-ras* gene was not significantly associated with overall survival (hazard ratio for death, 1.01;  $P=0.97$ ).

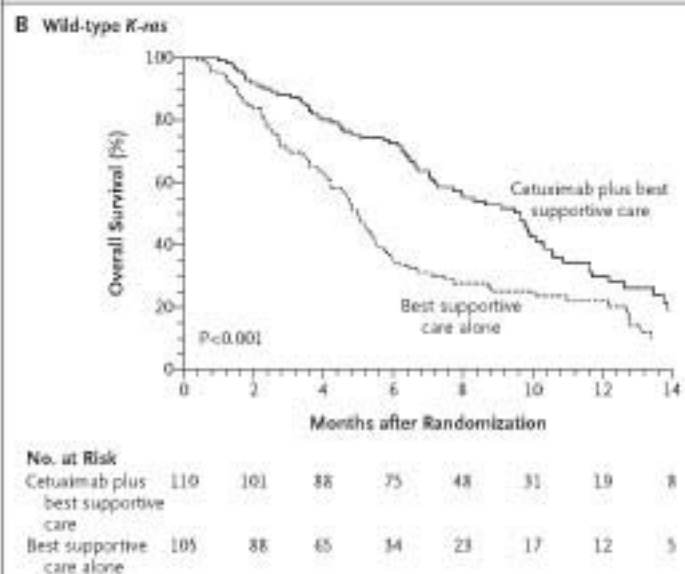
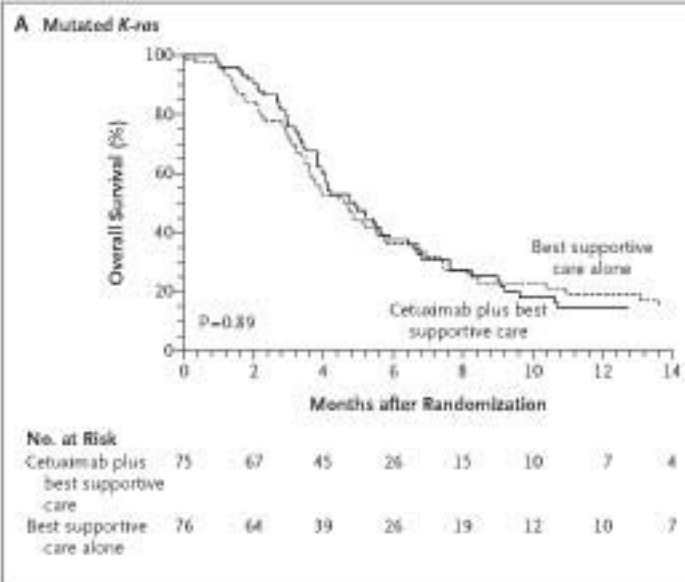
#### CONCLUSIONS

Patients with a colorectal tumor bearing mutated *K-ras* did not benefit from cetuximab, whereas patients with a tumor bearing wild-type *K-ras* did benefit from cetuximab. The mutation status of the *K-ras* gene had no influence on survival among patients treated with best supportive care alone. (ClinicalTrials.gov number, NCT00079066.)

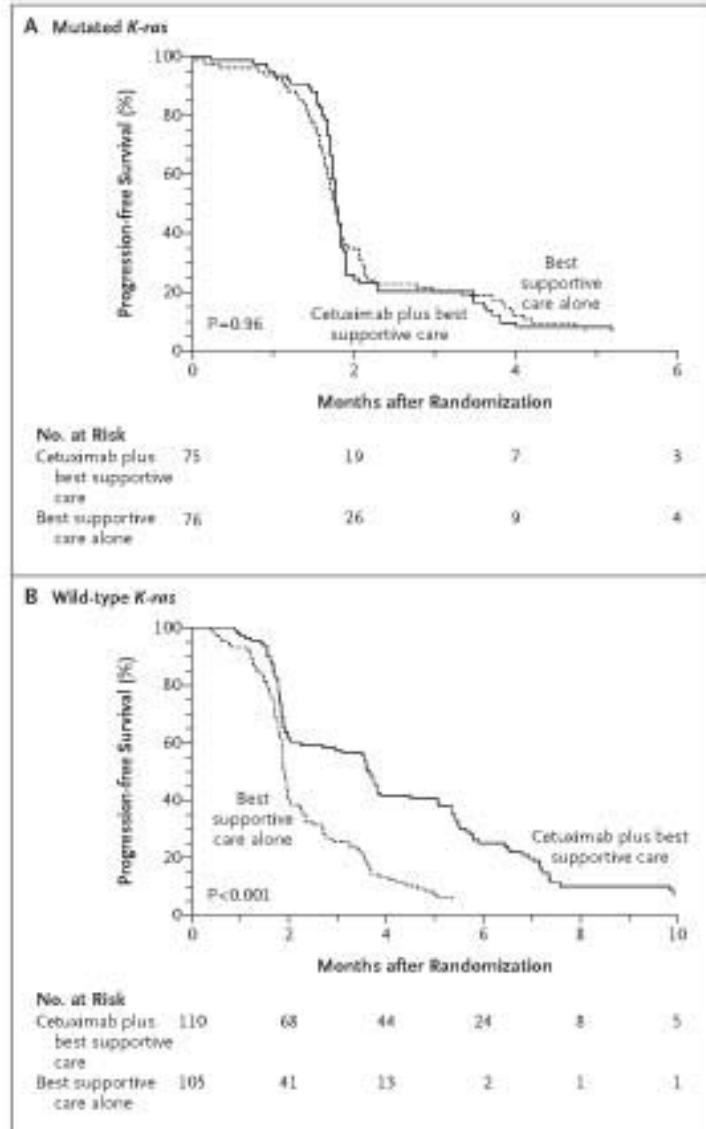
From Flinders Medical Centre and Flinders University, Adelaide, Australia (C.S.K.); Bristol-Myers Squibb Research and Development, Princeton, NJ (S.K.-F.); Ottawa Hospital Research Institute, University of Ottawa, Ottawa (D.J.); National Cancer Institute of Canada Clinical Trials Group, Kingston, ON (C.J.O., D.T., S.R., L.S.); Austin Health, Melbourne, Australia (N.C.T.); National Health and Medical Research Council Clinical Trials Centre, University of Sydney, Sydney (R.J.S.); Allan Blair Cancer Centre, Regina, SK, Canada (H.C.); Cabrini Hospital and Alfred Hospital, Melbourne, Australia (J.D.S.); Queen Elizabeth Hospital and University of Adelaide, Adelaide, Australia (T.J.P.); Cross Cancer Institute, Edmonton, AB, Canada (H.-J.A.); Bristol-Myers Squibb, Wallingford, CT (C.L.); Princess Margaret Hospital, Toronto (M.J.M.); and Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, Australia (J.R.Z.). Address reprint requests to Dr. Karapetis at the Department of Medical Oncology, Flinders Medical Centre, Flinders Dr. Bedford Park, SA 5042, Australia, or at c.karapetis@flinders.edu.au.

\*Other participants in the CO.17 trial from the National Cancer Institute of Canada Clinical Trials Group and the Australasian Gastro-Intestinal Trials Group are listed in the Supplementary Appendix, available with the full text of this article at [www.nejm.org](http://www.nejm.org).

N Engl J Med 2008;359:1757-65.  
Copyright © 2008 Massachusetts Medical Society.



**Figure 1. Kaplan–Meier Curves for Overall Survival According to Treatment.** Panel A shows results for patients with mutated *K-ras* tumors, and Panel B for patients with wild-type *K-ras* tumors. Cetuximab as compared with best supportive care alone was associated with improved overall survival among patients with wild-type *K-ras* tumors but not among those with mutated *K-ras* tumors. The difference in treatment effect according to mutation status was significant (test for interaction,  $P=0.01$ ).



**Figure 2. Kaplan–Meier Curves for Progression-free Survival According to Treatment.**

Panel A shows results for patients with mutated *K-ras* tumors, and Panel B for patients with wild-type *K-ras* tumors. Cetuximab as compared with best supportive care alone was associated with improved progression-free survival among patients with wild-type *K-ras* tumors but not among those with mutated *K-ras* tumors. The difference in treatment effect according to mutation status was significant (test for interaction,  $P<0.001$ ).



- Home
- Meetings
- Abstracts & Virtual Meeting
- Practice Resources
- Education & Training
- News
  - Press Center
  - ASCO News & Forum
  - Feature Articles
  - Podcasts
- Legislative & Regulatory
- Quality Care & Guidelines

- About ASCO
- ASCO Bookstore
- Careers in Oncology
- Downloads & Technology
- Foundation Grants & Awards
- Membership
- Research Policy
- State Affiliates



Home > News > Feature Articles

### ASCO Releases its First Provisional Clinical Opinion (PCO)

Patients with metastatic colorectal cancer who are candidates for anti-EGFR therapy should have their tumors tested for *KRAS* gene mutations, according to ASCO's first Provisional Clinical Opinion (PCO).

If a patient has a mutated form of the *KRAS* gene, the Society recommends *against* the use of anti-EGFR antibody therapy, based on recent studies indicating this treatment is only effective in patients with the normal (wild-type) form of the *KRAS* gene. It is estimated that 40% of patients with colon cancer have the *KRAS* mutation.

"Personalized medicine is the next frontier in cancer care," said Richard L. Schilsky, MD, ASCO President. "Using *KRAS* testing to guide colorectal cancer treatment is a prime example of where cancer care is heading."

"Basing cancer treatment on the unique genetic characteristics of the tumor or the individual with cancer will improve patient outcomes and help avoid unnecessary costs and side effects for patients who are unlikely to benefit," Dr. Schilsky added.

PCOs are intended to offer timely preliminary clinical direction to oncologists following the publication or presentation of potentially practice-changing data from major studies. ASCO's PCO on *KRAS* gene testing was given prior to the January 15-17, 2009 Gastrointestinal Cancers Symposium in San Francisco, California. The Symposium was co-sponsored by ASCO, the American Gastroenterological Association (AGA), the American Society for Radiation Oncology (ASTRO), and the Society of Surgical Oncology (SSO).

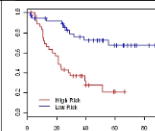
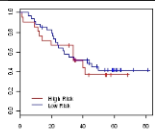
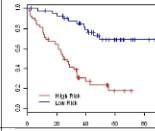
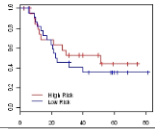
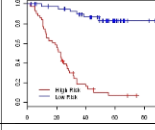
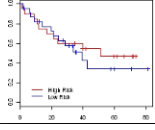
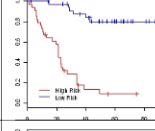
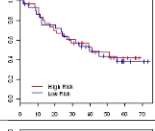
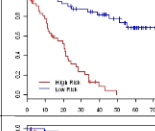
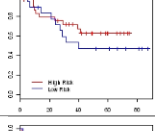
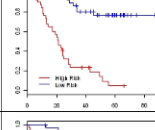
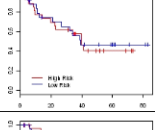
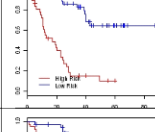
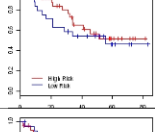
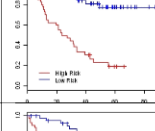
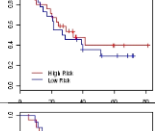
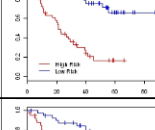
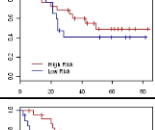
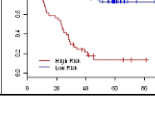
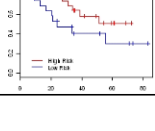
Among the 500 presentations was an important economic and scientific study that discussed the possibility of more than half a billion dollars in savings for the United States healthcare system. The study showed that routine testing for *KRAS* gene mutations in patients with metastatic colorectal cancer could save the U.S. health system up to \$604 million per year by identifying who would benefit from the drug cetuximab.

Information on the PCO is currently available on [ASCO.org](http://ASCO.org), and the entire report will be published in the February, 1 2009 issue of the *Journal of Clinical Oncology* (JCO).

# Biotechnology Has Forced Biostatistics to Focus on Prediction

- This has led to many interesting statistical developments
  - $p \gg n$  problems in which number of genes is much greater than the number of cases
- Growing pains in learning to address prediction problems
  - Many of the methods and much of the conventional wisdom of statistics are based on inference problems and are not applicable to prediction problems

- Goodness of fit is not a proper measure of predictive accuracy

Simulation	Training	Validation
1	 $p=7.0e-05$	 $p=0.70$
2	 $p=4.2e-07$	 $p=0.54$
3	 $p=2.4e-13$	 $p=0.60$
4	 $p=1.3e-10$	 $p=0.89$
5	 $p=1.8e-13$	 $p=0.36$
6	 $p=5.5e-11$	 $p=0.81$
7	 $p=3.2e-09$	 $p=0.46$
8	 $p=1.8e-07$	 $p=0.61$
9	 $p=1.1e-07$	 $p=0.49$
10	 $p=4.3e-09$	 $p=0.09$

# Prediction on Simulated Null Data

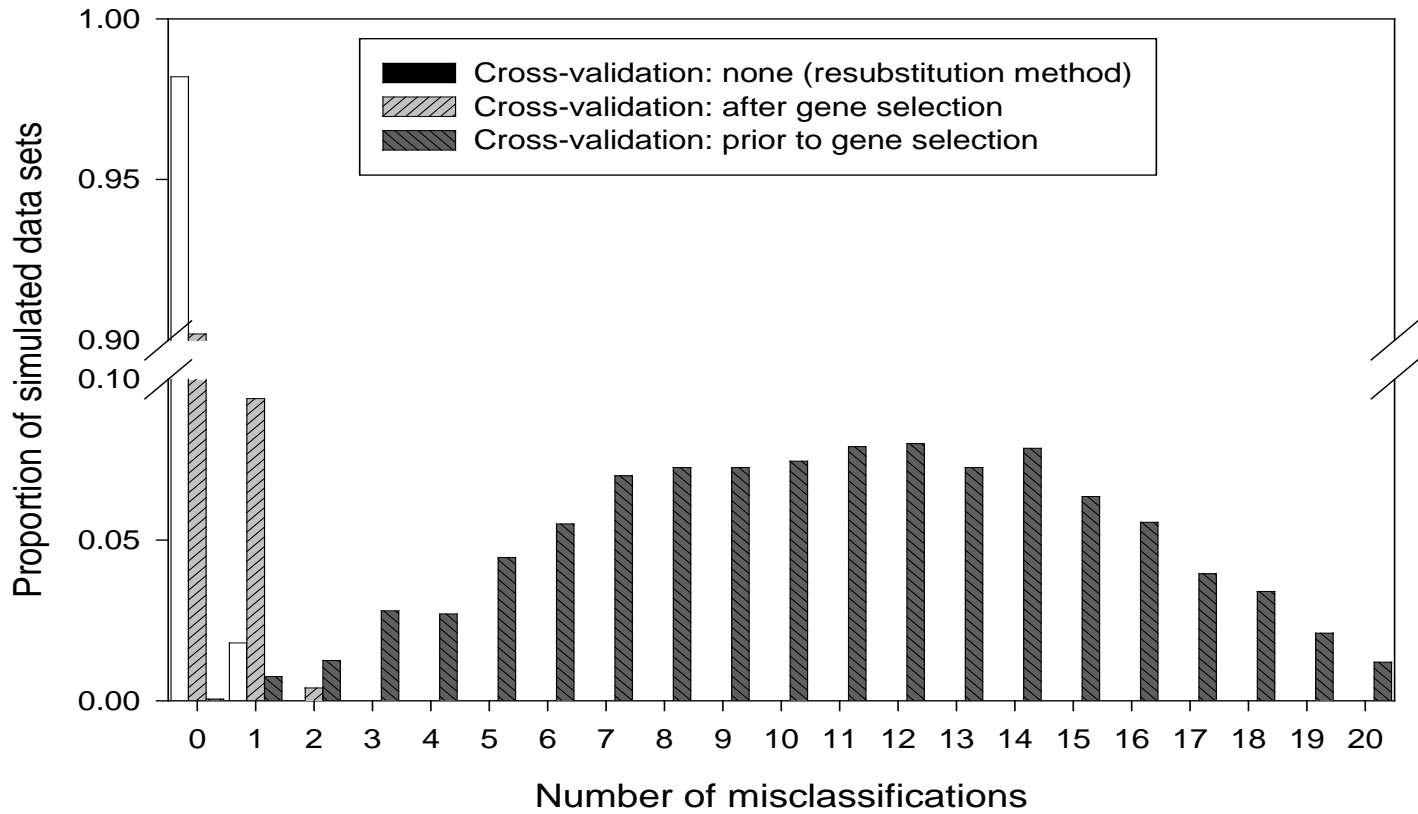
Simon et al. J Nat Cancer Inst 95:14, 2003

## Generation of Gene Expression Profiles

- 14 specimens ( $P_i$  is the expression profile for specimen  $i$ )
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- Can we distinguish between the first 7 specimens (Class 1) and the last 7 (Class 2)?

## Prediction Method

- Compound covariate predictor built from the log-ratios of the 10 most differentially expressed genes.



- **“Prediction is difficult; particularly the future.”**

# Cross Validation

- Cross-validation simulates the process of separately developing a model on one set of data and predicting for a test set of data not used in developing the model
- The cross-validated estimate of misclassification error is an estimate of the prediction error for the model developed by applying the specified algorithm to the full dataset



- 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.

# Permutation Distribution of Cross-validated Misclassification Rate of a Multivariate Classifier

Radmacher, McShane & Simon  
J Comp Biol 9:505, 2002

- Randomly permute class labels and repeat the entire cross-validation
- Re-do for all (or 1000) random permutations of class labels
- Permutation p value is fraction of random permutations that gave as few misclassifications as  $e$  in the real data

## Prediction of cancer outcome with microarrays: a multiple random validation strategy

Lancet 2005; 365: 488–92

See Comment page 454

**Biostatistics and Epidemiology Unit** (S Michiels, S Koscielny, PhD, C Hill PMD), **Functional Genomics Unit** (S Michiels), and **Inserm U605** (S Koscielny), Institut Gustave-Roussy, Villejuif, France

Correspondence to: Dr Serge Koscielny, Biostatistics and Epidemiology Unit, Institut Gustave Roussy, 39 rue Carle Desnoixes, 94805 Villejuif, France  
koscielny@igr.fr

Stefan Michiels, Serge Koscielny, Catherine Hill

### Summary

**Background** General studies of microarray gene-expression profiling have been undertaken to predict cancer outcome. Knowledge of this gene-expression profile or molecular signature should improve treatment of patients by allowing treatment to be tailored to the severity of the disease. We reanalysed data from the seven largest published studies that have attempted to predict prognosis of cancer patients on the basis of DNA microarray analysis.

**Methods** The standard strategy is to identify a molecular signature (ie, the subset of genes most differentially expressed in patients with different outcomes) in a training set of patients and to estimate the proportion of misclassifications with this signature on an independent validation set of patients. We expanded this strategy (based on unique training and validation sets) by using multiple random sets, to study the stability of the molecular signature and the proportion of misclassifications.

**Findings** The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. For all but one study, the proportion misclassified decreased as the number of patients in the training set increased. Because of inadequate validation, our chosen studies published overoptimistic results compared with those from our own analyses. Five of the seven studies did not classify patients better than chance.

**Interpretation** The prognostic value of published microarray results in cancer studies should be considered with caution. We advocate the use of validation by repeated random sampling.

### Introduction

The expression of several thousand genes can be studied simultaneously by use of DNA microarrays. These microarrays have been used in many specialities of medicine. In oncology, their use can identify genes with different expressions in tumours with different outcomes.<sup>1–3</sup> These gene-expression profiles or molecular signatures are expected to assist in the selection of optimum treatment strategies, by allowing therapy to be adapted to the severity of the disease.<sup>1,2</sup> Gene-expression profiling is already being used in clinical trials to define the population of patients with breast cancer who should receive chemotherapy. Such trials are being launched in Dutch academic centres and in the USA.<sup>4</sup>

A major challenge with DNA microarray technology is analysis of the massive data output, which needs to account for several sources of variability arising from the biological samples, hybridisation protocols, scanning, and image analysis.<sup>5</sup> Diverse approaches are used to classify patients on the basis of expression profiles: Fisher's linear discriminant analysis, nearest-neighbour prediction rule, and support vector machine, among others.<sup>6,7,8</sup> To estimate the accuracy of a classification method, the standard strategy is via a training-validation approach, in which a training set is used to identify the molecular signature and a validation set is used to estimate the proportion of misclassifications.

Leading scientific journals require investigators of DNA microarray research to deposit their data in an appropriate international database,<sup>9</sup> following a set of

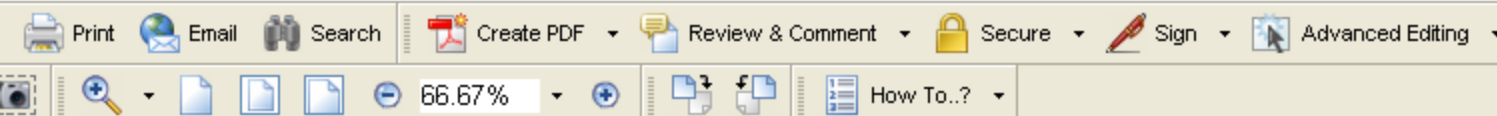
guidelines (Minimum Information About a Microarray Experiment<sup>10</sup>). This approach offers an opportunity to propose alternative analyses of these data. We have taken advantage of this opportunity to analyse different datasets from published studies of gene expression as a predictor of cancer outcome. We aimed to assess the extent to which the molecular signature depends on the constitution of the training set, and to study the distribution of misclassification rates across validation sets, by applying a multiple random training-validation strategy. We explored the relation between sample size and misclassification rates by varying the sample size in the training and validation sets.

### Methods

#### Data sources

All microarray studies of cancer prognosis published between January, 1995, and April, 2003, were reviewed in 2003 by Ntzani and Ioannidis.<sup>1</sup> From this review, we selected studies on survival-related outcomes (disease-free, event-free, or overall survival), which had included at least 60 patients (table). These studies used various classification methods: linear discriminant analysis, support vector machines, and prediction rules based on Cox's regression models. The sample size varied between 60 and 240 and the percentage of events between 14% and 58%.

Data were publicly available for seven studies<sup>1–7</sup> (webtable at <http://image.thelancet.com/extras/04art5032webtable.pdf>). We defined a binary clinical



## 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

- 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
- Misleading use of cluster analysis
  - 13/28 studies invalidly claimed that expression clusters based on differentially expressed genes could help distinguish clinical outcomes
- 50% of studies contained one or more major flaws

# Model Instability Does Not Mean Prediction Inaccuracy

- Validation of a predictive model means that the model predicts accurately for independent data
- Validation does not mean that the model is stable or that using the same algorithm on independent data will give a similar model
- With  $p > n$  and many genes with correlated expression, the classifier will not be stable.

## 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., Dmitry 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.

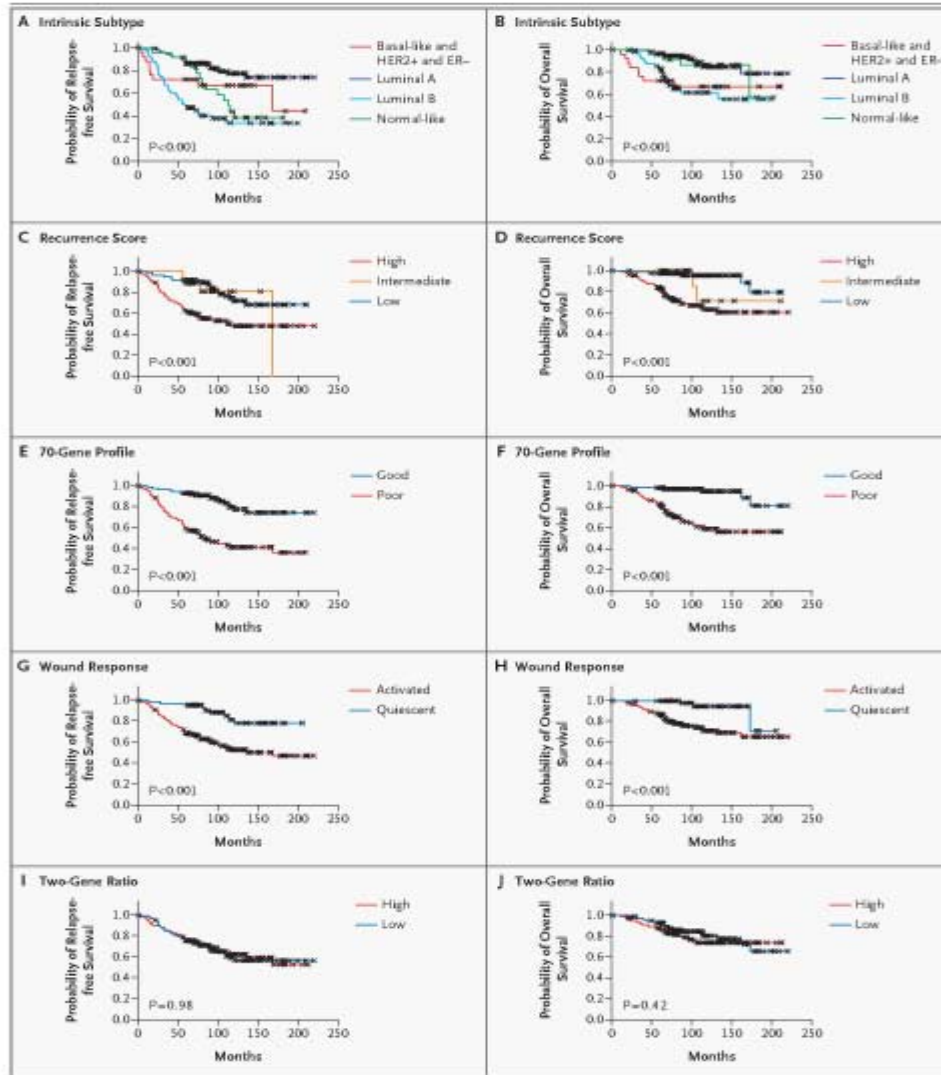
**CONCLUSIONS**

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.  
Copyright © 2006 Massachusetts Medical Society.

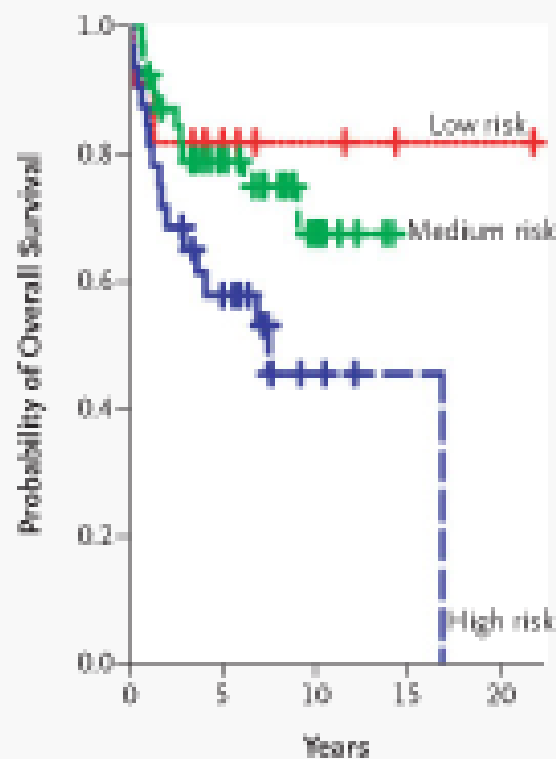




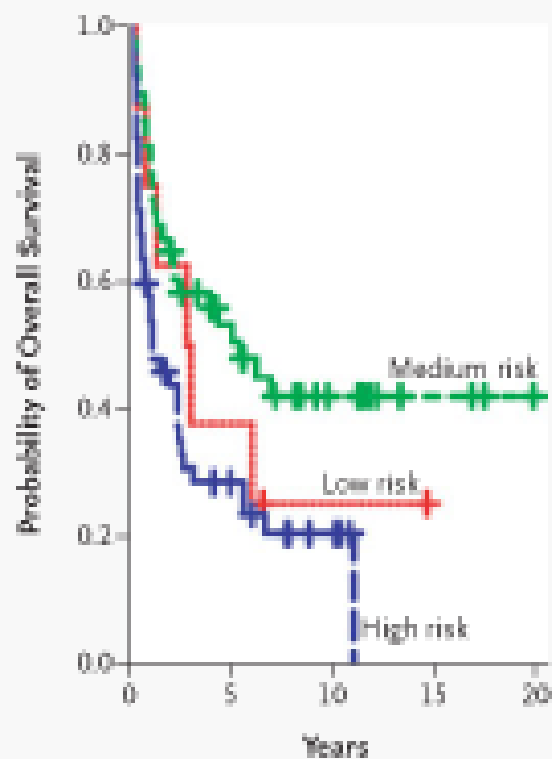
- **Odds ratios and hazards ratios are not proper measures of prediction accuracy**
- **Statistical significance of regression coefficients are not proper measures of predictive accuracy**

# Measures of Prognostic Value for Survival Data with a Test Set

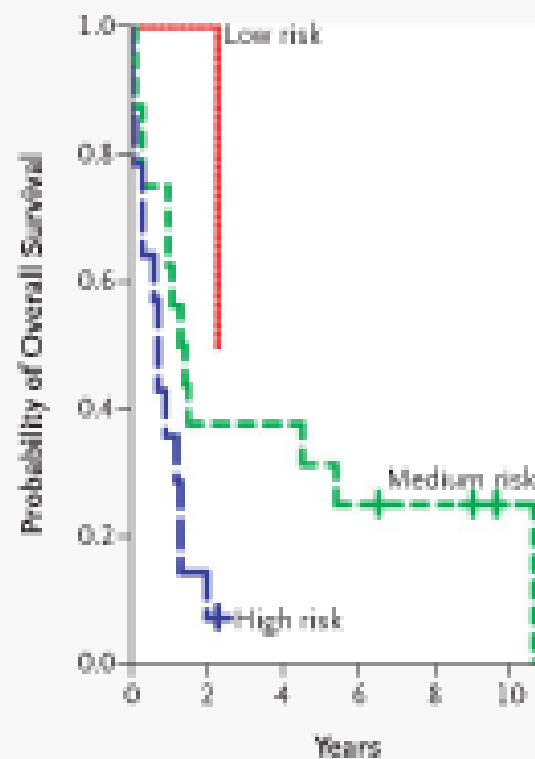
- A hazard ratio is a measure of association
  - Large values of HR may correspond to small improvement in prediction accuracy
- Kaplan-Meier curves on the test set for predicted risk groups within strata defined by standard prognostic variables provide more information about improvement in prediction accuracy
- Time dependent ROC curves on the test set within strata defined by standard prognostic factors can also be useful

**A Low IPI Risk Scores****No. at Risk**

Low	11	6	3	1	1
Medium	39	22	8	0	0
High	32	16	3	1	1

**B Medium IPI Risk Scores****No. at Risk**

Low	8	3	1	0	0
Medium	49	20	8	3	0
High	32	11	4	0	0

**C High IPI Risk Scores****No. at Risk**

Low	2	2	0	0	0	0
Medium	16	6	6	4	3	1
High	14	2	0	0	0	0

**Figure 4. The Six-Gene Model and the International Prognostic Index.**

The Kaplan–Meier estimates show overall survival for groups of patients with low-risk (Panel A), medium-risk (Panel B), and high-risk (Panel C) scores on the International Prognostic Index, as reported by Rosenwald et al.,<sup>6</sup> after subdivision into three groups (at low, medium, and high risk for death) on the basis of the six-gene model for prediction. According to log-likelihood estimates,  $P=0.01$ ,  $P=0.002$ , and  $P=0.16$  for the model based on a continuous variable applied to the low-risk, medium-risk, and high-risk groups, respectively, and  $P=0.02$ ,  $P=0.003$ , and  $P=0.01$ , respectively, for the model based on the three discrete groups shown in the figure.

# Does an Expression Profile Classifier Enable Improved Treatment Decisions Compared to Practice Standards?

- Not an issue of which variables are significant after adjusting for which others or which are *independent* predictors
- Requires focus on a defined medical indication
  - Selection of cases
  - Collection of covariate information
  - Analysis

# Is Accurate Prediction Possible For $p \gg n$ ?

- Yes, in many cases, but standard statistical methods for model building and evaluation are often not effective
  - Problem difficulty is often more important than algorithm used for variable selection or model used for classification
  - Often many models will predict adequately except complex models that over-fit the training data

- Standard regression methods are generally not useful for  $p > n$  problems
  - Standard methods may over-fit the data and lead to poor predictions
    - Estimating covariances, selecting interactions, transforming variables for improving goodness of fit, minimizing squared error often leads to over-fitting
    - Fisher LDA vs Diagonal LDA
    - With  $p > n$ , unless data is inconsistent, a linear model can always be found that classifies the training data perfectly

- $p > n$  prediction problems are not multiple testing problems
- The objective of prediction problems is accurate prediction, not controlling the false discovery rate
- Parameters that control feature selection in prediction problems are tuning parameters to be optimized for prediction accuracy

# Developing Predictive Models With $p > n$

- Gene selection is not a multiple testing problem
  - Predicting accurately
  - Testing hypotheses about which genes are correlated with outcome
  - Biological understanding
  - Are different problems which require different methods and resources



# Traditional Approach to Clinical Development a New Drug

- Small phase II trials to find primary sites where the drug appears active
- Phase III trials with broad eligibility to test the null hypothesis that a regimen containing the new drug is not better than the control treatment overall for all randomized patients
- If you reject  $H_0$  then treat all future patients satisfying the eligibility criteria with the new regimen, otherwise treat no such future patients with the new drug
- Perform subset hypotheses but don't believe them

# Traditional Clinical Trial Approaches

- Based on assumptions that
  - Qualitative treatment by subset interactions are unlikely
  - “Costs” of over-treatment are less than “costs” of under-treatment
- Neither of these assumptions is valid with most new molecularly targeted oncology drugs

# Traditional Clinical Trial Approaches

- Have protected us from false claims resulting from post-hoc data dredging not based on pre-defined biologically based hypotheses
- Have led to widespread over-treatment of patients with drugs to which many don't need and from which many don't benefit
- May have resulted in some false negative results

# Clinical Trials Should Be Science Based

- Cancers of a primary site may represent a heterogeneous group of diverse molecular diseases which vary fundamentally with regard to
  - their oncogenesis and pathogenesis
  - their responsiveness to specific drugs
- The established molecular heterogeneity of human cancer requires the use of new approaches to the development and evaluation of therapeutics

How Can We Develop New Drugs  
in a Manner More Consistent With  
Modern Tumor Biology and Obtain  
**Reliable** Information About What  
Regimens Work for What Kinds of  
Patients?

# 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

# Prospective Drug Development With a Companion Diagnostic

1. Develop a completely specified genomic classifier of the patients likely to benefit from a new drug
  - Larger phase II trials with evaluation of candidate markers
2. Establish analytical validity 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 that preserves the overall type-I error of the study.

Develop Predictor of Response to New Drug

```
graph TD; A[Develop Predictor of Response to New Drug] --> B[Patient Predicted Responsive]; A --> C[Patient Predicted Non-Responsive]; B --> D[New Drug]; B --> E[Control]; C --> F[Off Study];
```

Patient Predicted Responsive

Patient Predicted Non-Responsive

New Drug

Control

Off Study

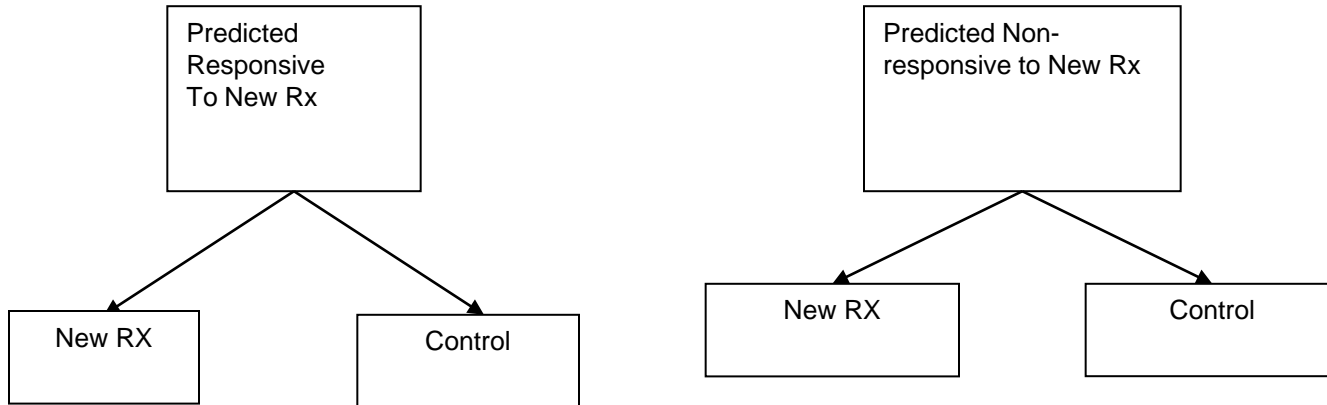


# Evaluating the Efficiency of Enrichment Design

- Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 10:6759-63, 2004; Correction and supplement 12:3229, 2006
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 24:329-339, 2005.
- R Simon. Using genomics in clinical trial design, *Clinical Cancer Research* 14:5984-93, 2008
- Reprints at <http://brb.nci.nih.gov>

# Developmental Strategy (II)

Develop Predictor of  
Response to New Rx



## Developmental Strategy (II)

- Do not use the diagnostic to restrict eligibility, but to structure a prospective analysis plan
- Having a prospective analysis plan is essential
- “Stratifying” (balancing) the randomization is useful to ensure that all randomized patients have tissue available but is not a substitute for a prospective analysis plan
- The purpose of the study is to evaluate the new treatment overall and for the pre-defined subsets; not 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

- R Simon. Using genomics in clinical trial design, *Clinical Cancer Research* 14:5984-93, 2008
- R Simon. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics, *Expert Opinion in Medical Diagnostics* 2:721-29, 2008

# Web Based Software for Designing RCT of Drug and Predictive Biomarker

- <http://brb.nci.nih.gov>

## *Biomarker Stratified Randomized Design*

Stratified design randomizes both marker positive and negative patients.

See references 73-75 in Technical Reports Section

- **Stratified Design with Prospective Analysis Plan and Binary Endpoint**
- **Stratified Design with Prospective Analysis Plan and Time-to-Event Endpoint**

© NIH, 2008

## Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect

Wenyu Jiang, Boris Freidlin, Richard Simon

### Background

Many molecularly targeted anticancer agents entering the definitive stage of clinical development benefit only a subset of treated patients. This may lead to missing effective agents by the traditional broad-eligibility randomized trials due to the dilution of the overall treatment effect. We propose a statistically rigorous biomarker-adaptive threshold phase III design for settings in which a putative biomarker to identify patients who are sensitive to the new agent is measured on a continuous or graded scale.

### Methods

The design combines a test for overall treatment effect in all randomly assigned patients with the establishment and validation of a cut point for a prespecified biomarker of the sensitive subpopulation. The performance of the biomarker-adaptive design, relative to a traditional design that ignores the biomarker, was evaluated in a simulation study. The biomarker-adaptive design was also used to analyze data from a prostate cancer trial.

### Results

In the simulation study, the biomarker-adaptive design preserved the power to detect the overall effect when the new treatment is broadly effective. When the proportion of sensitive patients as identified by the biomarker is low, the proposed design provided a substantial improvement in efficiency compared with the traditional trial design. Recommendations for sample size planning and implementation of the biomarker-adaptive design are provided.

### Conclusions

A statistically valid test for a biomarker-defined subset effect can be prospectively incorporated into a randomized phase III design without compromising the ability to detect an overall effect if the intervention is beneficial in a broad population.

J Natl Cancer Inst 2007;99:1-8

Human cancers are heterogeneous with regard to their molecular and genomic properties. Recent advances in biotechnology have resulted in a shift toward molecularly targeted anticancer agents. These new therapeutics are likely to benefit only a subset of the patients with a given cancer. Definitive testing of such targeted agents requires the identification of the appropriate "sensitive" population. When biomarkers to identify the patients who are likely to benefit from the new therapy are available, targeted clinical trials that restrict eligibility to sensitive patients should be used (1). However, reliable assays to identify sensitive patients are often unavailable. In the absence of a reliable biomarker, broad-eligibility clinical trials are used routinely. Most of these trials use a conventional design, in which the primary analysis is based on comparison of all randomly assigned patients. This often leads to the failure to recognize effective agents due to dilution of the treatment effect by the presence of the patients who do not benefit from the agent.

Retrospective analysis of trials with a conventional design can be used as an initial step in identifying biomarkers for the sensitive subpopulation. However, retrospectively identified biomarkers typically have to be validated in a confirmatory prospective randomized phase III clinical trial (2). This approach is inefficient and may considerably prolong clinical development.

Previously, we have proposed a design [adaptive signature design (3)] that combines a definitive test for treatment effect in a broad population with identification and validation of a genomic signature for the subset of sensitive patients if the broad population test is negative. The adaptive signature design was developed for high-dimensional data such as gene expression microarrays, where only a few unknown genes among thousands assayed may be relevant and where a classifier (signature) to identify sensitive patients is not available. The design incorporates both the identification and the validation of a pharmacogenomic signature for sensitive patients.

Often, preliminary information on a biomarker to identify the sensitive subset of patients is available but an appropriate cutoff

**Affiliation of authors:** Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD.

**Correspondence to:** Boris Freidlin, PhD, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, EPN 8122, National Cancer Institute, Bethesda, MD 20892 (e-mail: freidlinb@ctep.nci.nih.gov).

See "Notes" following "References."

DOI: 10.1093/jnci/dkj022

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oxfordjournals.org.

## Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients

Boris Freidlin and Richard Simon

**Abstract Purpose:** A new generation of molecularly targeted agents is entering the definitive stage of clinical evaluation. Many of these drugs benefit only a subset of treated patients and may be overlooked by the traditional, broad-eligibility approach to randomized clinical trials. Thus, there is a need for development of novel statistical methodology for rapid evaluation of these agents. **Experimental Design:** We propose a new adaptive design for randomized clinical trials of targeted agents in settings where an assay or signature that identifies sensitive patients is not available at the outset of the study. The design combines prospective development of a gene expression-based classifier to select sensitive patients with a properly powered test for overall effect. **Results:** Performance of the adaptive design, relative to the more traditional design, is evaluated in a simulation study. It is shown that when the proportion of patients sensitive to the new drug is low, the adaptive design substantially reduces the chance of false rejection of effective new treatments. When the new treatment is broadly effective, the adaptive design has power to detect the overall effect similar to the traditional design. Formulas are provided to determine the situations in which the new design is advantageous. **Conclusion:** Development of a gene expression-based classifier to identify the subset of sensitive patients can be prospectively incorporated into a randomized phase III design without compromising the ability to detect an overall effect.

Developments in tumor biology have resulted in shift toward molecularly targeted drugs (1–3). Most human tumor types are heterogeneous with regard to molecular pathogenesis, genomic signatures, and phenotypic properties. As a result, only a subset of the patients with a given cancer is likely to benefit from a targeted agent (4). This complicates all stages of clinical development, especially randomized phase III trials (5, 6). In some cases, predictive assays that can accurately identify patients who are likely to benefit from the new therapy have been developed. Then, targeted randomized designs that restrict eligibility to patients with sensitive tumors should be used (7). However, reliable assays to select sensitive patients are often not available (8, 9). Consequently, traditional randomized clinical trials with broad eligibility criteria are routinely used to evaluate such agents. This is generally inefficient and may lead to missing effective agents.

Genomic technologies, such as microarrays and single nucleotide polymorphism genotyping, are powerful tools that hold a great potential for identifying patients who are likely to benefit from a targeted agent (10, 11). However, due to the large number of genes available for analysis, interpretation of these data is complicated. Separation of reliable evidence from the random patterns inherent in high-dimensional data requires specialized statistical methodology that is prospectively incorporated in the trial design. Practical implementation of such designs has been lagging. In particular, analysis of microarray data from phase III randomized studies is usually considered secondary to the primary overall comparison of all eligible patients. Many analyses are not explicitly written into protocols and done retrospectively, mainly as "hypothesis-generating" tools.

We propose a new adaptive design for randomized clinical trials of molecularly targeted agents in settings where an assay or signature that identifies sensitive patients is not available. Our approach includes three components: (a) a statistically valid identification, based on the first stage of the trial, of the subset of patients who are most likely to benefit from the new agent; (b) a properly powered test of overall treatment effect at the end of the trial using all randomized patients; and (c) a test of treatment effect for the subset identified in the first stage, but using only patients randomized in the remainder of the trial. The components are prospectively incorporated into a single phase III randomized clinical trial with the overall false-positive error rate controlled at a prespecified level.

**Authors' Affiliation:** Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland

Received 3/18/05; revised 7/18/05; accepted 8/4/05. 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.

**Requests for reprints:** Boris Freidlin, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 8130 Executive Boulevard, EPN 8122, MSC 7434, Bethesda, MD 20892-7434. Phone: 301-402-0640; Fax: 301-402-0560; E-mail: freidlinb@ctep.nci.nih.gov. doi:10.1158/1078-0432.CCR.05-0105

# Multiple Biomarker Design

A Generalization of the Biomarker Adaptive Threshold Design

- Have identified  $K$  candidate binary classifiers  $B_1, \dots, B_K$  thought to be predictive of patients likely to benefit from  $T$  relative to  $C$
- RCT comparing new treatment  $T$  to control  $C$
- Eligibility not restricted by candidate classifiers
- Let the  $B_0$  classifier classify all patients positive



- Test T vs C restricted to patients positive for  $B_k$  for  $k=0,1,\dots,K$ 
  - Let  $S(B_k)$  be a measure of treatment effect in patients positive for  $B_k$
  - Let  $S^* = \max\{S(B_k)\}$  ,  $k^* = \operatorname{argmax}\{S(B_k)\}$
  - $S^*$  is the largest treatment effect observed
  - $k^*$  is the marker that identifies the patients where the largest treatment effect is observed

- For a global test of significance
  - Randomly permute the treatment labels and repeat the process of computing  $S^*$  for the shuffled data
  - Repeat this to generate the distribution of  $S^*$  under the null hypothesis that there is no treatment effect for any subset of patients
  - The statistical significance level is the area in the tail of the null distribution beyond the value of  $S^*$  obtained for the un-shuffled data
  - If the data value of  $S^*$  is significant at 0.05 level, then claim effectiveness of T for patients positive for marker  $k^*$

- Repeating the analysis for bootstrap samples of cases provides
  - an estimate of the stability of  $k^*$  (the indication)

**Cross-Validated  
Adaptive Signature Design**  
(submitted for publication)

**Wenyu Jiang, Boris Freidlin,  
Richard Simon**

# Cross-Validated Adaptive Signature Design End of Trial Analysis

- Compare T to C for **all patients** at significance level  $\alpha_{\text{overall}}$ 
  - If overall  $H_0$  is rejected, then claim effectiveness of T for eligible patients
  - Otherwise

# Otherwise

- Partition the full data set into  $K$  parts
- Form a training set by omitting one of the  $K$  parts. The omitted part is the test set
  - Using the training set, develop a predictive classifier of the subset of patients who benefit preferentially from the new treatment  $T$  compared to control  $C$  using the methods developed for the ASD
  - Classify the patients in the test set as sensitive (classifier +) or insensitive (classifier -)
- Repeat this procedure  $K$  times, leaving out a different part each time
  - After this is completed, all patients in the full dataset are classified as sensitive or insensitive

- Compare T to C for sensitive patients by computing a test statistic S e.g. the difference in response proportions or log-rank statistic (for survival)
- Generate the null distribution of S by permuting the treatment labels and repeating the entire K-fold cross-validation procedure
- Perform test at significance level 0.05 -  $\alpha_{\text{overall}}$
- If  $H_0$  is rejected, claim effectiveness of T for subset defined by classifier
  - The sensitive subset is determined by developing a classifier using the full dataset

70% Response to T in Sensitive Patients  
25% Response to T Otherwise  
25% Response to C  
20% Patients Sensitive

	ASD	CV-ASD
Overall 0.05 Test	0.486	0.503
Overall 0.04 Test	0.452	0.471
Sensitive Subset 0.01 Test	0.207	0.588
Overall Power	0.525	0.731



# Prediction Based Analysis of Clinical Trials

- Using cross-validation we can evaluate our methods for analysis of clinical trials, including complex subset analysis algorithms, in terms of their effect on improving patient outcome via informing therapeutic decision making

# Conclusions

- Personalized Oncology is Here Today and Rapidly Advancing
  - Key information is in tumor genome
  - Read-out is about biology of the tumor, not susceptibility for possible disease or adverse effects

# Conclusions

- Some of the conventional wisdom about statistical analysis of clinical trials is not applicable to trials dealing with co-development of drugs and diagnostics
  - e.g. subset analysis if the overall results are not significant or if an interaction test is not significant

# Conclusions

- Co-development of drugs and companion diagnostics increases the complexity of drug development
  - It does not make drug development simpler, cheaper and quicker
  - But it may make development more successful and it has great potential value for patients and for the economics of health care

# Conclusions

- Biotechnology is forcing statisticians to address problems of prediction
- Many existing statistical paradigms for model development and validation are not effective for  $p > n$  problems
- New approaches to the design and analysis of RCTs that both test an overall  $H_0$  and inform treatment decisions for individual patients are needed

# Acknowledgements

- NCI Biometric Research Branch
  - Kevin Dobbin
  - Boris Freidlin
  - Sally Hunsberger
  - Wenyu Jiang
  - Aboubakar Maitournam
  - Michael Radmacher
  - Yingdong Zhao
- BRB-ArrayTools Development Team