Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Pitfalls in the Development and Validation of Prognostic & Predictive Biomarker Classifiers

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

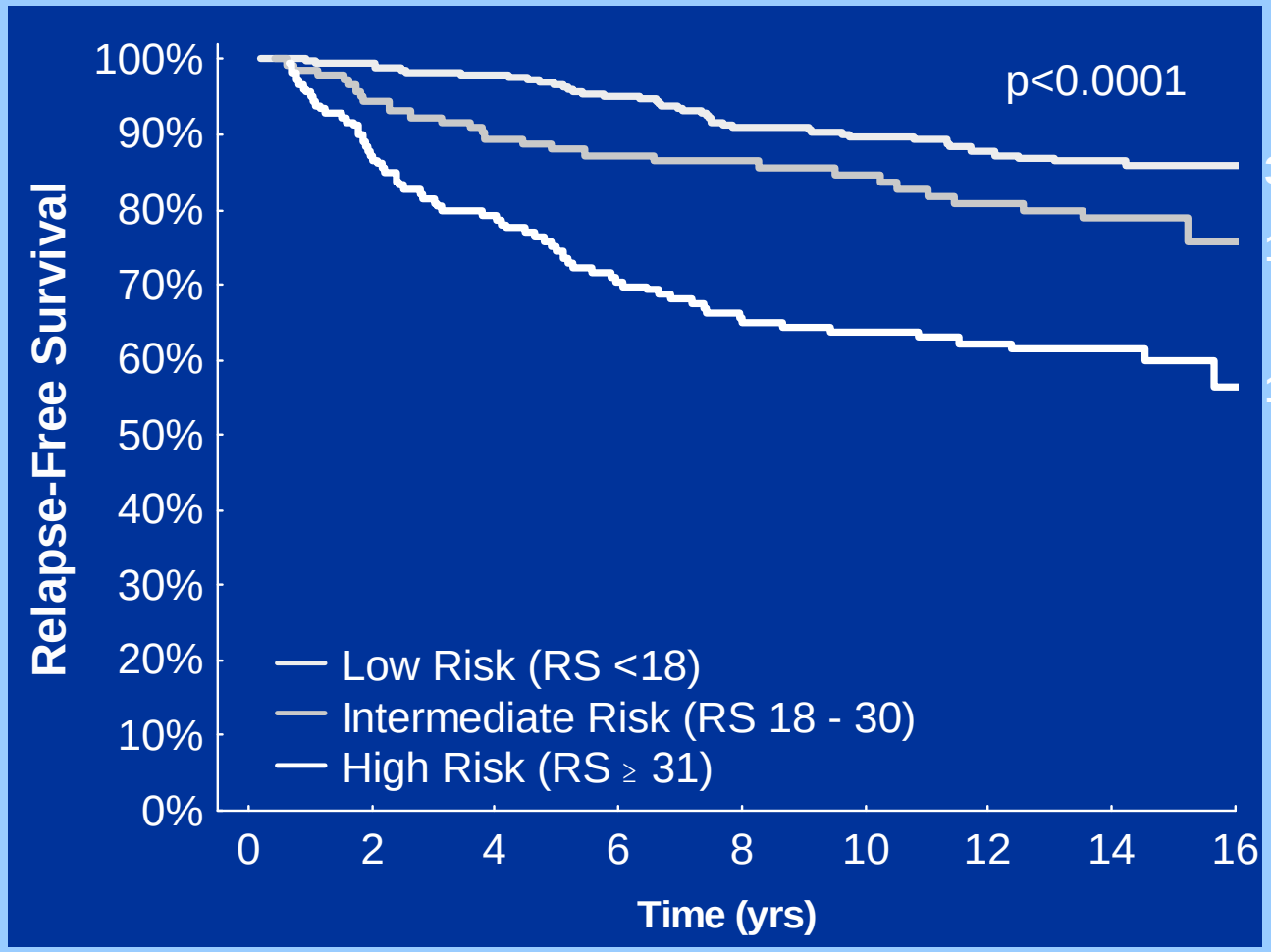
Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Several Kinds of Biomarkers

- Surrogate endpoints
  - A measurement made on a patient before, during and after treatment to determine whether the treatment is working
- Prognostic biomarkers
  - Measured before treatment to indicate long-term outcome for patients untreated or receiving standard treatment
- Predictive biomarkers
  - Measured before treatment to identify who will benefit from a particular treatment

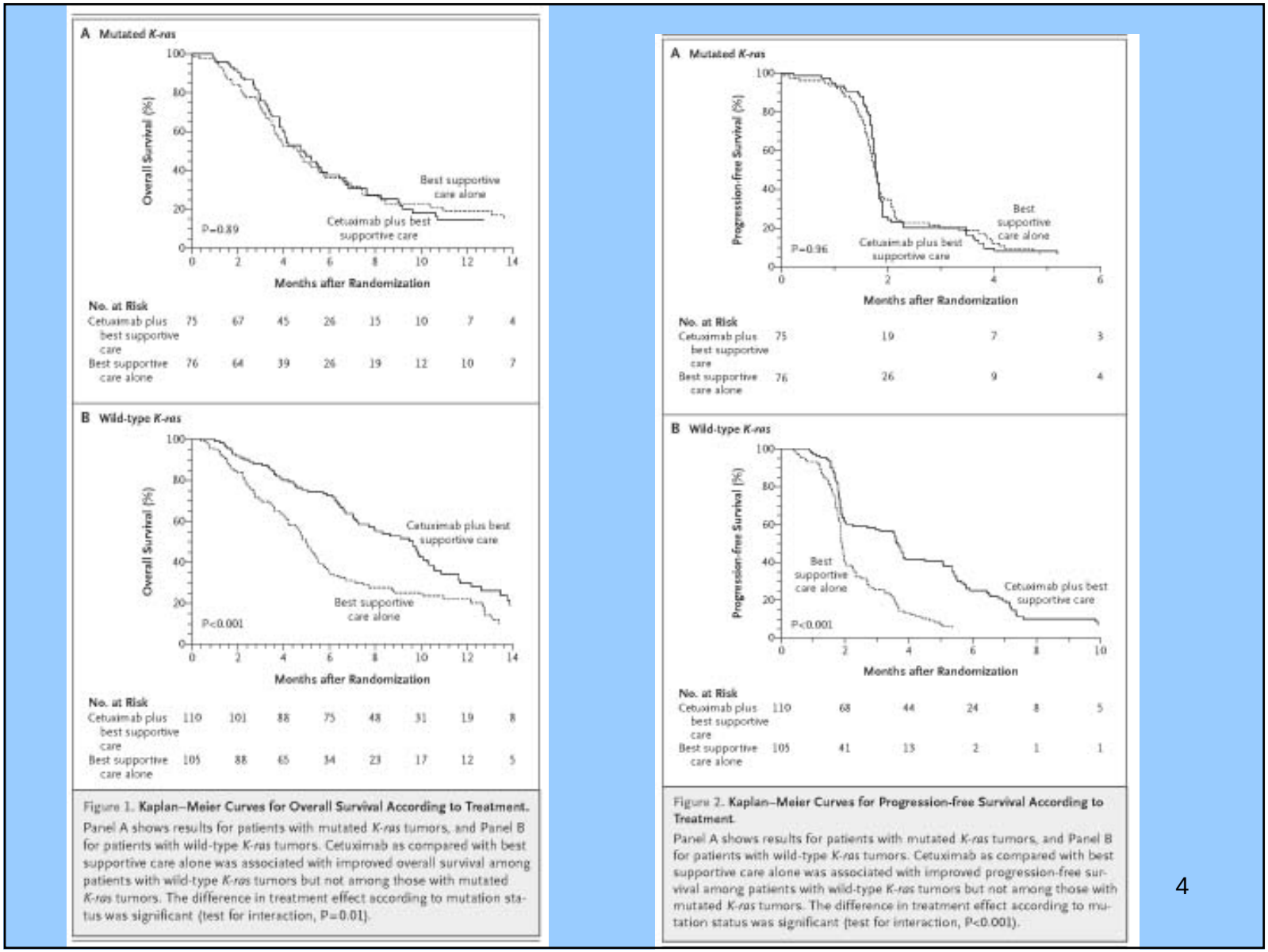
Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

B-14 Results—Relapse-Free Survival



Paik et al, SABCS 2003

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute



Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Prognostic and Predictive Biomarkers in Oncology

- Single gene or protein measurement
  - e.g. HER2 protein staining 2+ or 3+
  - HER2 amplification
  - KRAS mutation
- Scalar index or classifier that summarizes contributions of multiple genes/proteins
  - Empirically determined based on genome-wide correlating gene expression to patient outcome after treatment

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Prognostic & Predictive Biomarkers

- Many 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, and enhance their chance of receiving a drug that helps them
  - Help control medical costs
  - Improve the success rate of clinical drug development

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Prognostic Factors in Oncology

- Most prognostic factors are not used because they are not therapeutically relevant
- Most prognostic factor studies do not have a clear medical objective
  - They use a convenience sample of patients for whom tissue is available.
  - Generally the patients are too heterogeneous to support therapeutically relevant conclusions

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Key Features of OncotypeDx Development

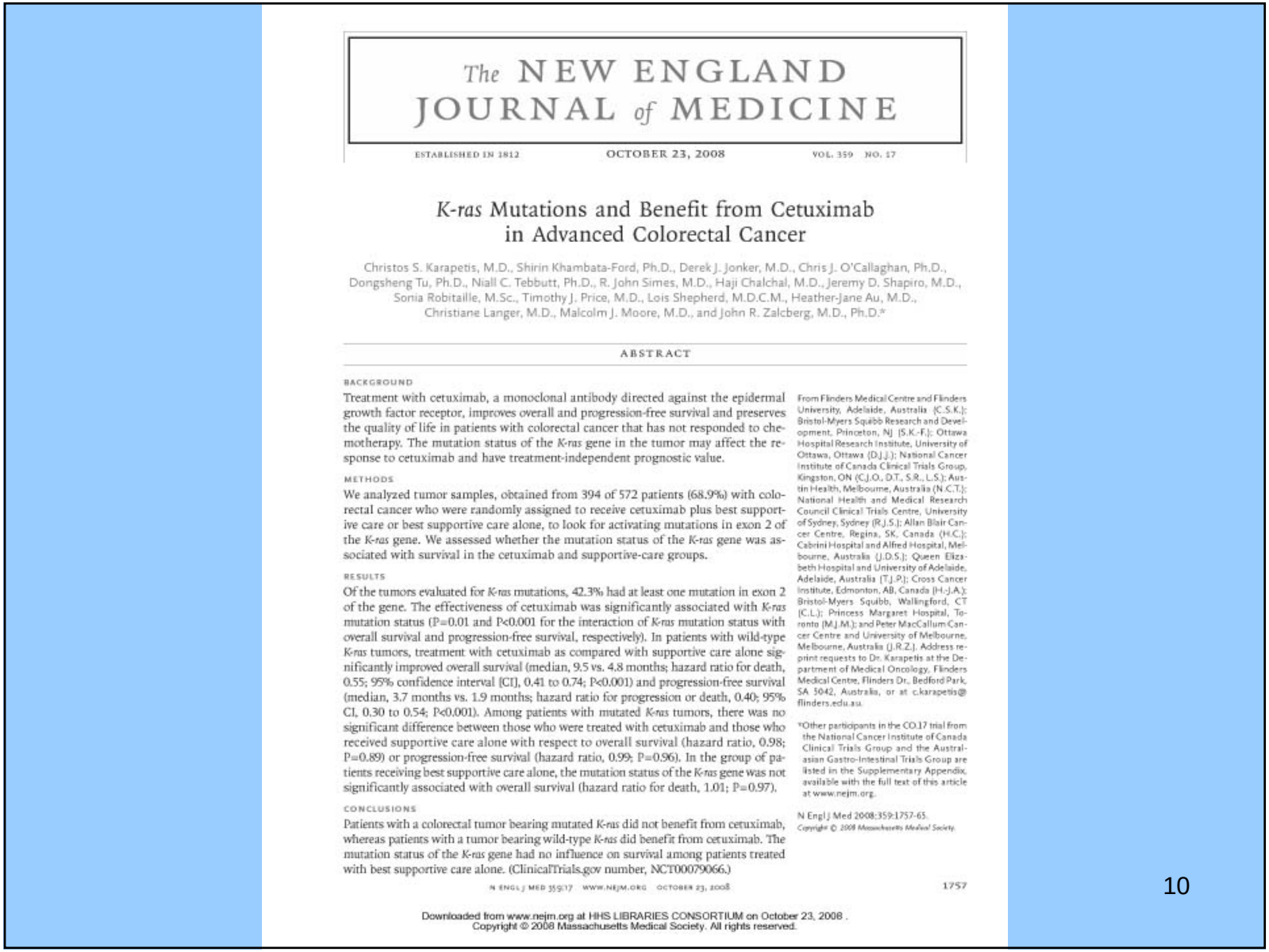
- Identification of important therapeutic decision context
- Prognostic marker development was based on patients with node negative ER positive breast cancer receiving tamoxifen as only systemic treatment
  - Use of patients in NSABP clinical trials
- Staged development and validation
  - Separation of data used for test development from data used for test validation
- Development of robust assay with rigorous analytical validation
  - 21 gene RTPCR assay for FFPE tissue
  - Quality assurance by single reference laboratory operation

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Predictive Biomarkers

- In the past often studied as un-focused post-hoc subset analyses of RCTs.
  - Numerous subsets examined
  - Same data used to define subsets for analysis and for comparing treatments within subsets
  - No control of type I error

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 OCTOBER 23, 2008 VOL. 359 NO. 17

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

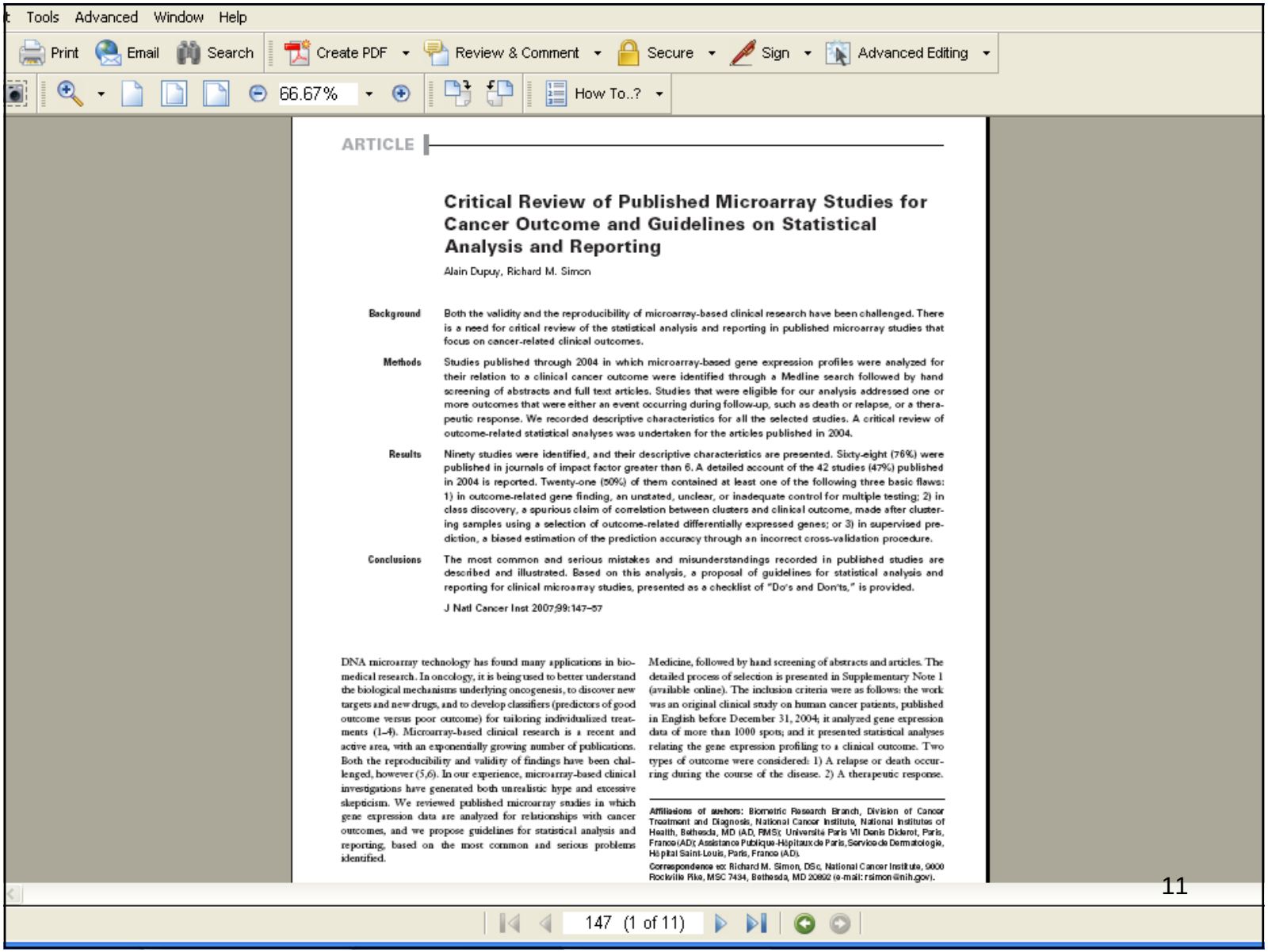
Patients with a colorectal tumor bearing mutated *K-ras* did not benefit from cetuximab, whereas patients with a tumor bearing wild-type *K-ras* did benefit from cetuximab. The mutation status of the *K-ras* gene had no influence on survival among patients treated with best supportive care alone. (ClinicalTrials.gov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

N Engl J Med 2008;359:1757-65.

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute



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.

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).

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Publications Reviewed

- Searched Medline
- Hand screening of abstracts & papers
- Original study on human cancer patients
- Published in English before December 31, 2004
- Analyzed gene expression of more than 1000 probes
- Related gene expression to clinical outcome

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

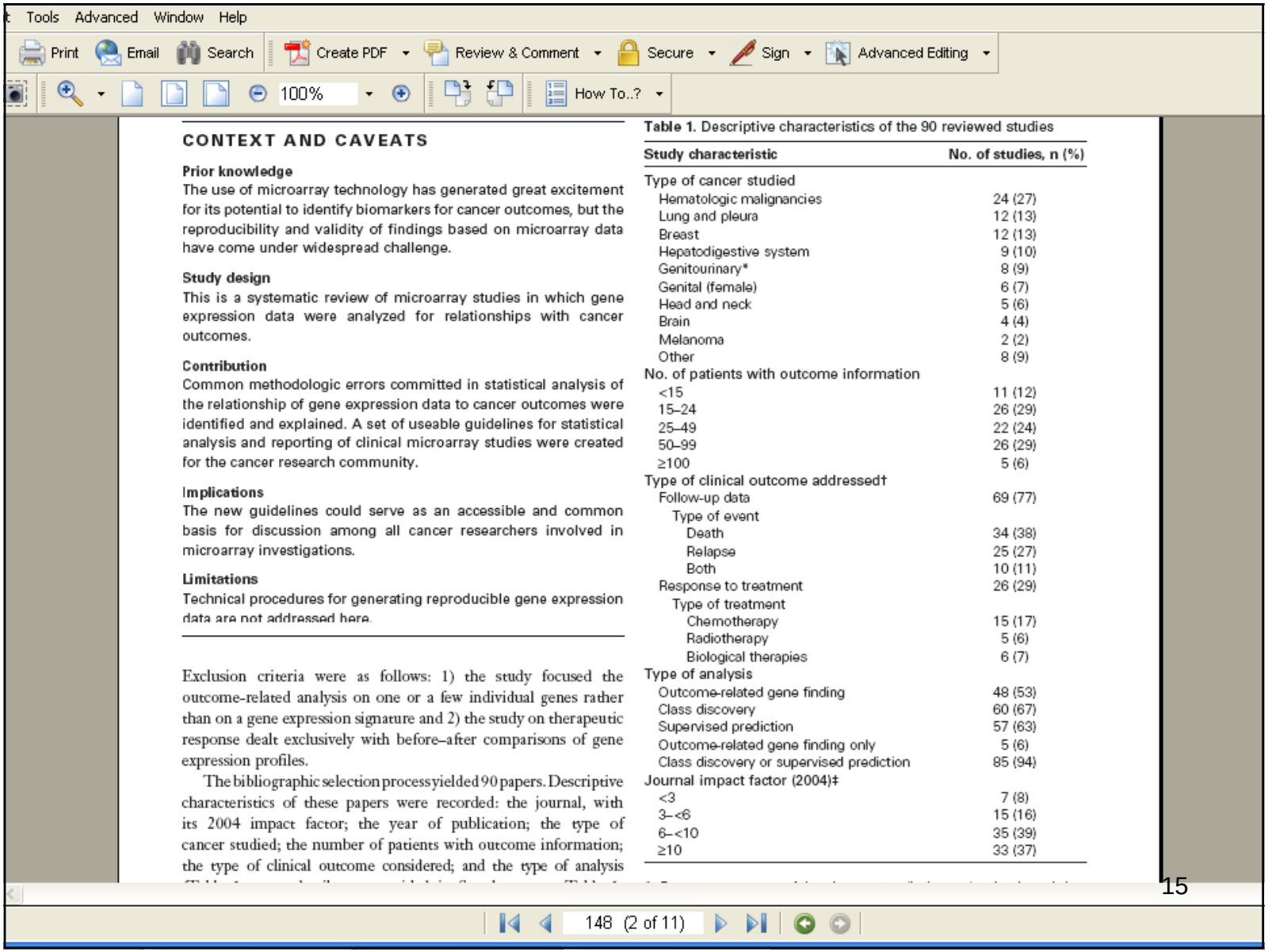
Types of Clinical Outcome

- Survival or disease-free survival
- Response to therapy

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

- 90 publications identified that met criteria
  - Abstracted information for all 90
- Performed detailed review of statistical analysis for the 42 papers published in 2004

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute



Tools Advanced Window Help

Print Email Search Create PDF Review & Comment Secure Sign Advanced Editing

100% How To..?

CONTEXT AND CAVEATS

Prior knowledge
The use of microarray technology has generated great excitement for its potential to identify biomarkers for cancer outcomes, but the reproducibility and validity of findings based on microarray data have come under widespread challenge.

Study design
This is a systematic review of microarray studies in which gene expression data were analyzed for relationships with cancer outcomes.

Contribution
Common methodologic errors committed in statistical analysis of the relationship of gene expression data to cancer outcomes were identified and explained. A set of useable guidelines for statistical analysis and reporting of clinical microarray studies were created for the cancer research community.

Implications
The new guidelines could serve as an accessible and common basis for discussion among all cancer researchers involved in microarray investigations.

Limitations
Technical procedures for generating reproducible gene expression data are not addressed here.

Exclusion criteria were as follows: 1) the study focused the outcome-related analysis on one or a few individual genes rather than on a gene expression signature and 2) the study on therapeutic response dealt exclusively with before–after comparisons of gene expression profiles.

The bibliographic selection process yielded 90 papers. Descriptive characteristics of these papers were recorded: the journal, with its 2004 impact factor; the year of publication; the type of cancer studied; the number of patients with outcome information; the type of clinical outcome considered; and the type of analysis

Table 1. Descriptive characteristics of the 90 reviewed studies

Study characteristic	No. of studies, n (%)
Type of cancer studied	
Hematologic malignancies	24 (27)
Lung and pleura	12 (13)
Breast	12 (13)
Hepatodigestive system	9 (10)
Genitourinary*	8 (9)
Genital (female)	6 (7)
Head and neck	5 (6)
Brain	4 (4)
Melanoma	2 (2)
Other	8 (9)
No. of patients with outcome information	
<15	11 (12)
15–24	26 (29)
25–49	22 (24)
50–99	26 (29)
≥100	5 (6)
Type of clinical outcome addressed†	
Follow-up data	69 (77)
Type of event	
Death	34 (38)
Relapse	25 (27)
Both	10 (11)
Response to treatment	26 (29)
Type of treatment	
Chemotherapy	15 (17)
Radiotherapy	5 (6)
Biological therapies	6 (7)
Type of analysis	
Outcome-related gene finding	48 (53)
Class discovery	60 (67)
Supervised prediction	57 (63)
Outcome-related gene finding only	5 (6)
Class discovery or supervised prediction	85 (94)
Journal impact factor (2004)‡	
<3	7 (8)
3–<6	15 (16)
6–<10	35 (39)
≥10	33 (37)

148 (2 of 11)

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

- Good microarray studies have clear objectives, but not generally gene specific mechanistic hypotheses
- Case selection and analysis methods should be tailored to study objectives

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Good Microarray Studies Have Clear Objectives

- Class Comparison (Gene Finding)
  - 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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Class Comparison and Class Prediction

- Not clustering problems
  - Global similarity measures generally used for clustering arrays may not distinguish classes
  - Don't control multiplicity or for distinguishing data used for classifier development from data used for classifier evaluation
- Supervised methods

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Do's & Don'ts of Gene Finding

- Don't use only fold-changes between groups to select genes
- Don't use a .05 significance threshold to select the differentially expressed genes
  - .05*10,000=500 false positives per 10,000 genes tested
- Do use a method to control the number of false positive differentially expressed genes or the false discovery rate

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Analysis Strategies for Gene Finding

- Compare classes on a gene by gene basis using statistical tests
  - Control for the large number of tests performed
  - e.g. use 0.0001 threshold of significance

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Analysis Strategies for Gene Finding

- Select the most differentially expressed genes in a manner that limits the false discovery rate to a specified level (e.g. 10%)

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

	Not rejected	Rejected	Total
True null hypotheses	890	10 False discoveries	900
False null hypotheses	10	90 True discoveries	100
		100	1000

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Methods for Controlling the False Discovery Rate

- Benjamini-Hochberg
- SAM (Tocher et al.)
- Multivariate permutation test (Korn et al.)

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Components of Class Prediction

- Gene selection
  - Which genes will be included in the model
- Model type selection
  - e.g. Diagonal linear discriminant analysis, Nearest-Neighbor, …
- Fitting parameters for model
  - Weights (regression coefficients)
  - Cut-points
  - Tuning parameters

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Myth

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

- Comparative studies indicate that
  - Standard statistical model development methods often over-fit the data and result in poor predictions
  - Complex “artificial intelligence” methods are often improperly evaluated and perform no better than simpler methods on realistic problems
  - Predictive classifiers designed to avoid over-fitting generally perform as well or better than other methods

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

- Predictive classifiers designed to avoid over-fitting generally perform as well or better than other methods
  - Gene selection based on univariate correlation with outcome
  - Model type linear or nearest neighbor classifiers

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Linear Classifiers for Two Classes

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

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

F = features (genes) included in model

w_i = weight for i'th feature

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

29

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Linear Classifiers for Two Classes

- Fisher linear discriminant analysis
- Diagonal linear discriminant analysis (DLDA) assumes features are uncorrelated
- Compound covariate predictor (Radmacher)
- Golub's weighted voting method
- Support vector machines with inner product kernel

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Advantages of Simple Linear Classifiers

- Do not over-fit data
  - Incorporate influence of multiple variables without attempting to select the best small subset of variables
  - Do not attempt to model the multivariate interactions among the predictors and outcome

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Nearest Neighbor Classifiers

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Validating 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 vs prediction accuracy
- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Validating a Classifier

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

The NEW ENGLAND JOURNAL of MEDICINE

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.

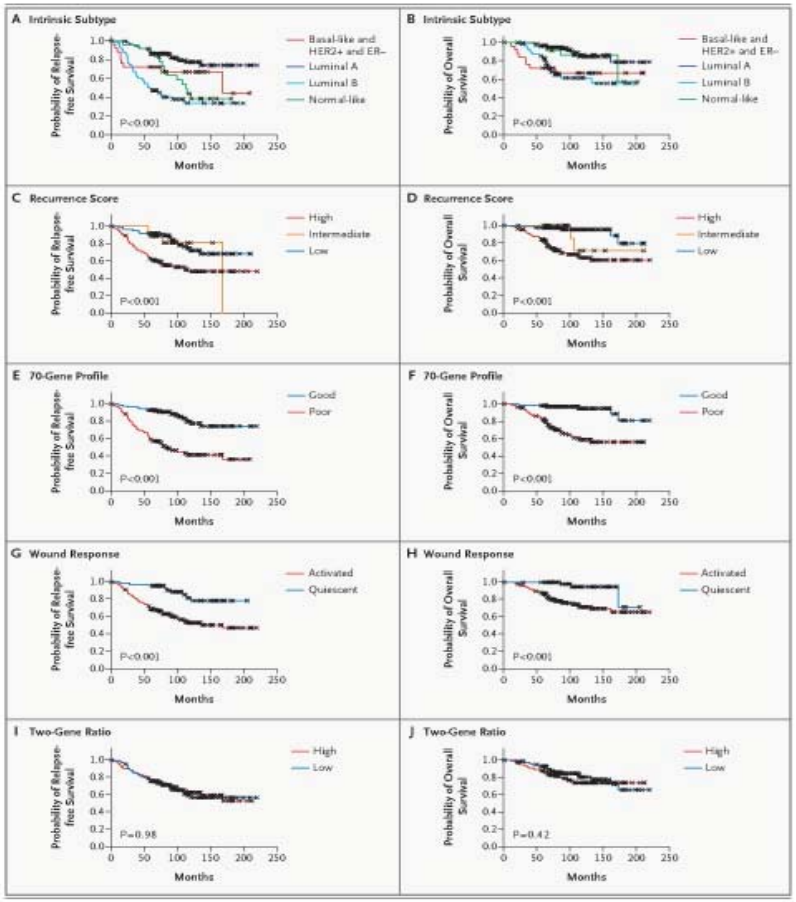
From the Departments of Genetics (C.F., D.S.O., C.M.P.), Statistics and Operations Research (A.B.N.), and Pathology and Laboratory Medicine (C.M.P.), University of North Carolina at Chapel Hill and Lineberger Comprehensive Cancer Center, Chapel Hill; and the Divisions of Diagnostic Oncology (L.W., B.W., 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.

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

CONCORDANCE AMONG GENE-EXPRESSION–BASED PREDICTORS FOR BREAST CANCER



Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Types of Validation for Prognostic and Predictive Biomarkers

- Analytical validation
  - Pre-analytical and post-analytical robustness
- Clinical validation
  - Does the biomarker predict what it's supposed to predict for independent data
- Clinical utility
  - Does use of the biomarker result in patient benefit

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Split-Sample Evaluation

- Used for Rosenwald et al. study of prognosis in DLBL lymphoma.
  - 200 cases training-set
  - 100 cases test-set

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Leave-one-out Cross Validation

- Leave-one-out 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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Leave-one-out Cross Validation

- Omit sample 1
  - Develop multivariate classifier from scratch on training set with sample 1 omitted
  - Predict class for sample 1 and record whether prediction is correct

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Leave-one-out Cross Validation

- Repeat analysis for training sets with each single sample omitted one at a time
- e = number of misclassifications determined by cross-validation
- Subdivide e for estimation of sensitivity and specificity

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Prediction on Simulated Null Data

Generation of Gene Expression Profiles

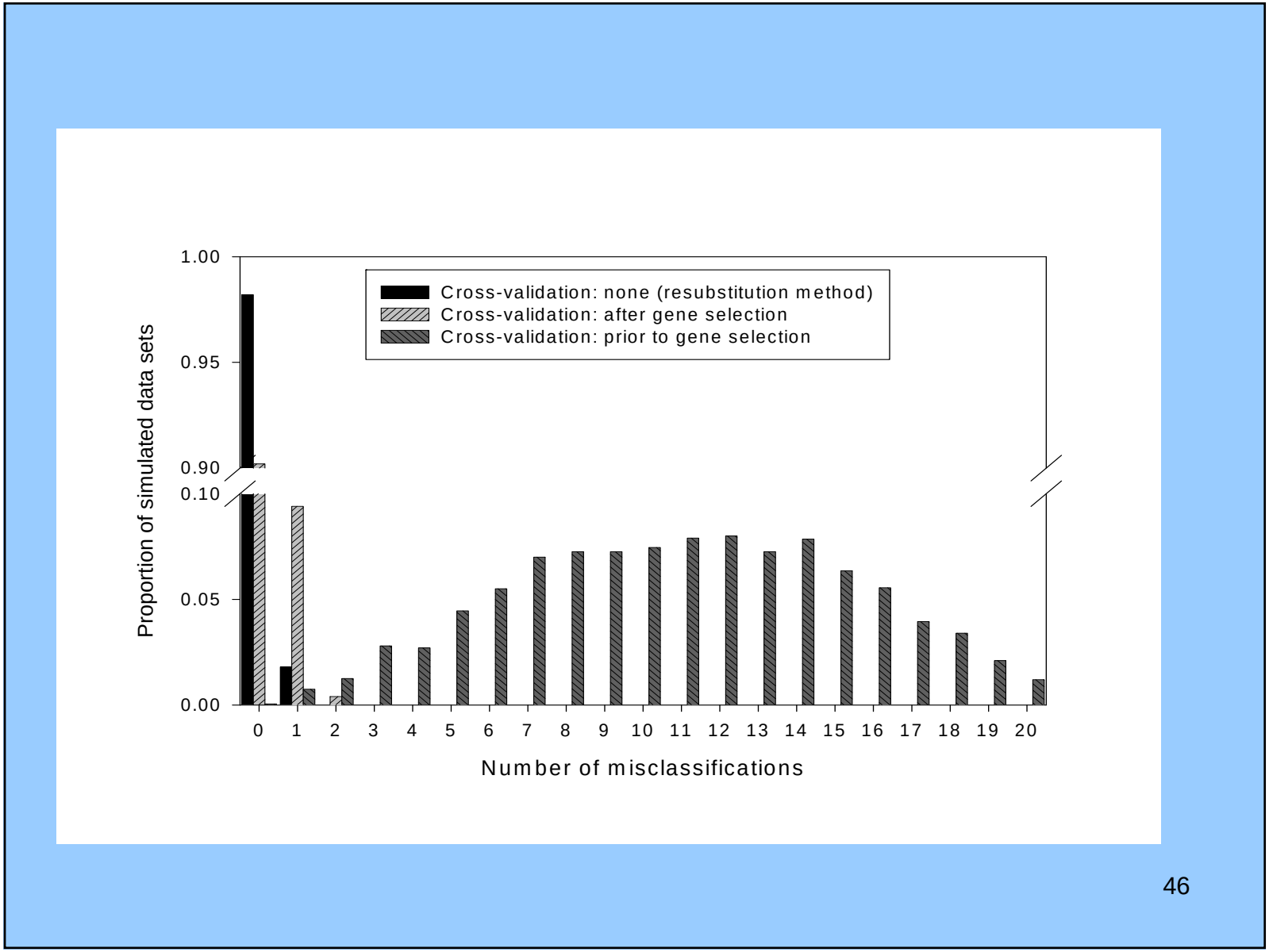
- 14 specimens (P_i is the expression profile for specimen i)
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- Can we distinguish between the first 7 specimens (Class 1) and the last 7 (Class 2)?

Prediction Method

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

45

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute



Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Myth

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Comparison of Internal Validation Methods

Molinaro, Pfiffer & Simon

- For small sample sizes, LOOCV is much less biased than split-sample validation
- For small sample sizes, LOOCV is preferable to 10-fold, 5-fold cross-validation or repeated k-fold versions
- For moderate sample sizes, 10-fold is preferable to LOOCV
- Some claims for bootstrap resampling for estimating prediction error are not valid for p>>n problems

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Do's & Don'ts
Supervised Prediction

- Do frame a therapeutically relevant question and select a homogeneous set of patients
- Don't violate the fundamental principle of classifier validation, i.e. ***no*** preliminary use of the test samples

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Do's & Don'ts
Separate Test Set

- Don't use any information from the test set in developing the classifier
- Do access the test set once and only for testing the fully specified classifier developed with the training data

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Do's & Don'ts
Cross Validation

- Don't use the same set of features for all iterations
- Do report error estimates for all classification methods tried, not just the one with the smallest error estimate
- Don't consider that retaining a small separate test set adds value to a correctly cross-validated estimate of accuracy
- Do report the fully specified classifier with its parameters

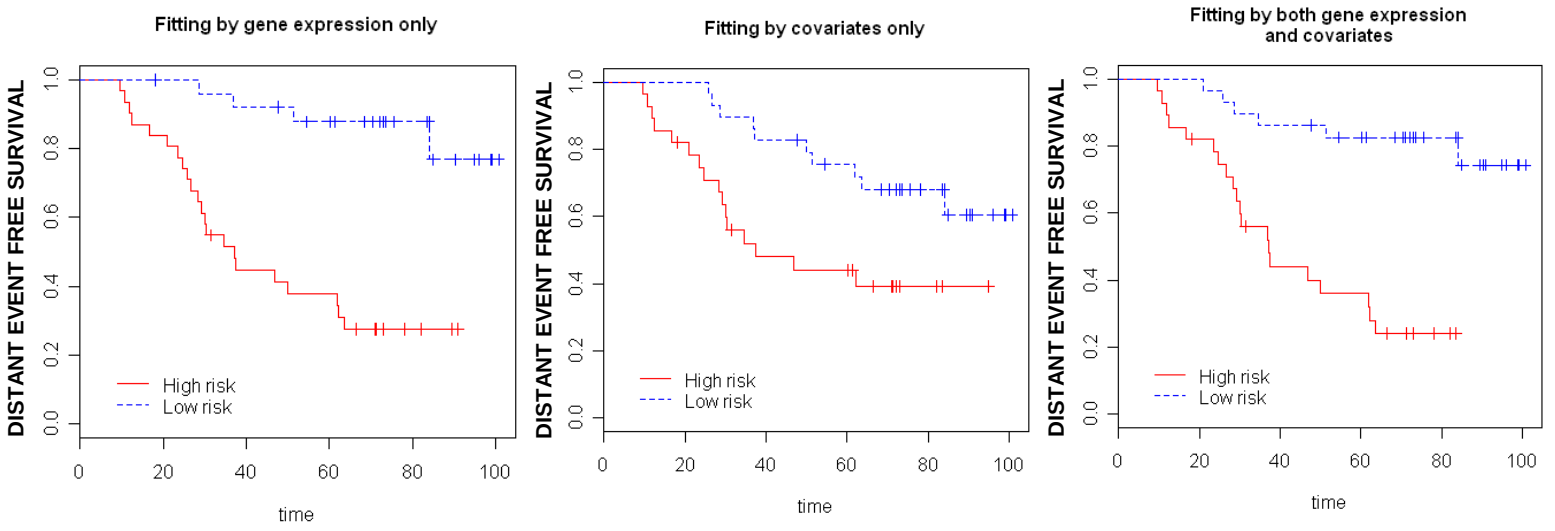
Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Myth

- For analyzing right censored data to develop predictive classifiers it is necessary to make the data binary

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Cross-validated Kaplan-Meier curves for risk groups using 50th percentile cut-off



GENE MODEL

COVARIATES MODEL

COMBINED MODEL

54

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Does an Expression Profile Classifier Predict More Accurately Than Standard Prognostic Variables?

- Not an issue of which variables are significant after adjusting for which others or which are *independent* predictors
  - Predictive accuracy and inference are different

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Cluster Analysis of Samples

- For discovering unanticipated structure and subsets of tissues

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Cluster Analysis is Subjective

- Cluster algorithms always produce clusters
- Different clustering algorithms may find different structure using the same data.

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Class Comparison and Class Prediction

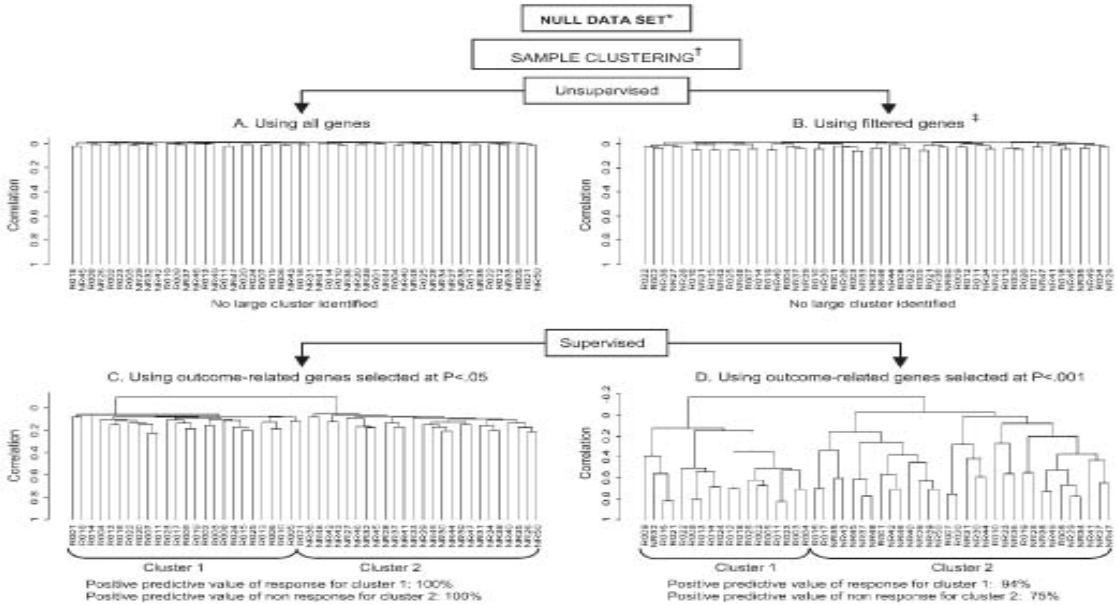
- Not clustering problems
  - Global similarity measures generally used for clustering arrays may not distinguish classes
  - Don't control multiplicity or for distinguishing data used for classifier development from data used for classifier evaluation
- Supervised methods

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Do's & Don'ts
Cluster Analysis

- Don't use "supervised" cluster analysis based on genes selected as differentially expressed among classes

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute



E. Simulation study of clustering null data sets* (sample clustering using partition around medoids with 2 clusters)

Simulation	Criteria for selecting genes used for sample clustering	Number of simulated data sets	Number of genes Mean (SD)	Positive predictive value of response for cluster 1 Mean (SD)	Positive predictive value of nonresponse for cluster 2 Mean (SD)
1	None	500	10000 (0)	55.7% (5.1)	55.3% (4.5)
2	Filtering	500	6785 (47)	55.2% (4.1)	55.4% (4.5)
3	DEG 0.05§	500	498 (21)	99.7% (1.3)	99.8% (1.3)
4	DEG 0.001§	500	9.6 (3.0)	86.8% (1.1)	86.9% (1.0)

Fig. 1. A–D) Dendrograms of hierarchical sample clustering of a null dataset* in four situations according to the type of gene selection (mentioned above each dendrogram) performed before clustering. In cases shown in **A** and **B**, unsupervised clustering failed to identify clusters correlated with the clinical outcome. This is the expected result for clustering a null dataset with a randomly allocated outcome. In cases shown in **C** and **D**, clustering was priorily supervised by using the response information to select the genes. The two identified clusters correlate convincingly with the clinical outcome, demonstrating that clustering was actually outcome driven through the prior selection of outcome-related differentially expressed genes. A claim for having discovered clinically meaningful clusters by correlating cluster and clinical outcome would therefore be spurious. **E)** Results of a simulation study using 500 different null datasets. Simulations 1–4 reproduced the type of gene selection used in **A–D**, respectively. In simulations 1 and 2, when clustering is unsupervised, no correlation between clusters and outcome categories is evidenced. In simulations 3 and 4, when clustering has been supervised by selecting outcome-related differentially expressed genes, a spurious correlation between cluster and outcome is evidenced. *The null dataset incorporates 10000 genes and 50 samples. Gene expression data values originate from a normal distribution (mean = 0; standard deviation = .05). The outcome is binary: response or nonresponse to a treatment. Half of the samples were randomly allocated to being from a responder, the other half to being from a nonresponder. †Hierarchical clustering using centered Pearson correlation metric and average linkage. ‡A gene was filtered out if less than 20% of its expression data values had at least 1.5-fold change in either direction from the gene's median value; § = Differentially expressed genes (DEG) using a .05 or .001 *P* value threshold for *t* test between outcome-defined classes.

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

BRB-ArrayTools

- Architect – R Simon
- Developer – Emmes Corporation

- Contains wide range of analysis tools selected by R Simon
- Designed for use by biomedical scientists
- Imports data from all gene expression and copy-number platforms
  - Automated import of data from NCBI Gene Express Omnibus
- Highly computationally efficient
- Extensive annotations for identified genes
- Integrated analysis of expression data, copy number data, pathway data and data other biological data

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

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
  - With clinical covariates
  - Cross-validated K-M curves
- Predict quantitative trait
  - LARS, LASSO
- Feature selection options
  - Univariate t/F statistic
  - Hierarchical random variance model
  - Restricted by fold effect
  - Univariate classification power
  - Recursive feature elimination
  - Top-scoring pairs
- Validation methods
  - Split-sample
  - LOOCV
  - Repeated k-fold CV
  - .632+ bootstrap
- Permutational statistical significance

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

BRB-ArrayTools

July 2008

- 8934 Registered users
- 68 Countries
- 616 Citations
- 19,628 hits/month to website

- Registered users
  - 4655 in US
    - 898 at NIH
      - 387 at NCI
    - 2994 US EDU
    - 1161 US Gov (non NIH)
  - 4655 Non US

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Countries With Most BRB ArrayTools Registered Users

- Germany 292
- France 289
- Canada 287
- UK 278
- Italy 250
- China 241
- Netherlands 240
- Taiwan 222
- Korea 192
- Japan 187
- Spain 168
- Australia 155
- India 139
- Belgium 103
- New Zeland 63
- Brazil 54
- Singapore 53
- Denmark 52
- Sweden 50
- Israel 45

Richard Simon, D.Sc.
Biometric Research Branch
Division of Cancer Treatment and Diagnosis, National Cancer Institute

Acknowledgements

- Alain Dupuy
- Annette Molinaro
- Ruth Pfiffer
- Michael Radmacher
- Yingdong Zhao
- BRB-ArrayTools Development Team