Development and Evaluation of Predictive Classifiers

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

Class prediction ≠ Class comparison

- Class comparison (gene finding)
  - Hierarchical model for variance
    - GW Wright & R Simon, Bioinformatics 19:2448-55, 2003
  - Multivariate permutation test based
    - SAM
    - Korn et al. JSPI 124:379-98, 2004
- Class prediction
  - Predictive accuracy

$$PCC(n) \geq \Phi\left\{\frac{m(1-\beta)(\delta/\sigma)}{\sqrt{\lambda}\sqrt{m(1-\beta)(p-m)\alpha}}\right\}$$

KK Dobbin, R Simon. Sample size planning for developing classifiers using high dimensional DNA microarray data, Biostatistics (in press)

Validation of Predictive Classifier Does Not Involve

- Measuring overlap of gene sets used in classifier developed from independent data
- Statistical significance of gene expression levels or summary signatures in multivariate analysis
- Confirmation of gene expression measurements on other platforms
- Demonstrating that the classifier or any of its components are “validated biomarkers of disease status”

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.

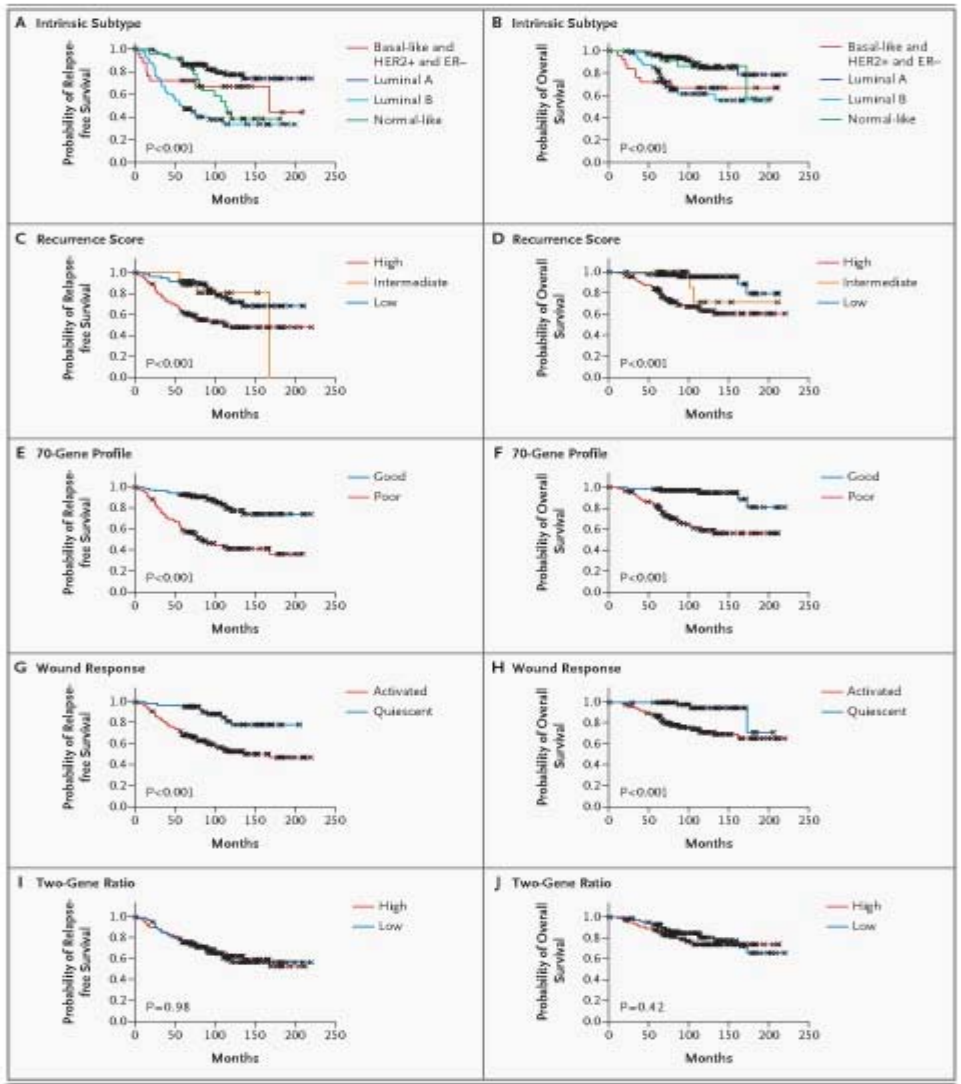
CONCLUSIONS

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.



Metrics that Matter

- Predictive accuracy
- Reproducibility of classification for individual patients
- Medical utility

Developmental Studies vs Validation Studies

- Developmental studies are exploratory
- FDA should not regulate classifier development
- Developmental studies should incorporate an internal estimate of predictive accuracy
  - Split sample
  - Cross-validation or bootstrap

Myth

- Split sample validation is superior to LOOCV or 10-fold CV for estimating prediction error

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Prediction Error Estimation: A Comparison of Resampling Methods

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

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

ABSTRACT

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

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

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

Contact: annette.molinaro@yale.edu

1 INTRODUCTION

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

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

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

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

*to whom correspondence should be addressed

- Both split-sample validation and cross-validation represent *internal validation*

Limitations to Internal Validation

- Sample handling and assay conduct are performed under controlled conditions that do not incorporate real world sources of variability
- Study on archived tissue may confound tissue handling or assay performance with outcome (class)
- Cases may be from a single institution
- Developmental studies are generally small
- Study does not establish reproducibility of classification for individual patients
- Predictive accuracy is generally not clinical utility

Predictive Accuracy ≠ Medical Utility

- Prognostic factors are generally not useful unless they have therapeutic relevance
- Predictive factors can be of great importance
  - Who benefits from a particular treatment
- Most developmental studies use a convenience sample of patients for whom tissue is available. Generally the patients are too heterogeneous to support therapeutically relevant conclusions
  - Focus on patients in a single clinical trial

“Validation” Is Worse than Meaningless
Except as
“Fit for Purpose”

Predictive Classifier
Fit For Purpose

- Adoption of a classifier to restrict the use of a treatment in wide use should be based on demonstrating that use of the classifier leads to better clinical outcome
- In new drug development, the role of a classifier is to select a target population for treatment
  - The focus should be on evaluating the new drug, not on “validating” the classifier

Validation Study
Node negative Breast Cancer

- Prospective study design
- Samples collected and assayed from patients with node negative ER+ breast cancer who will receive TAM
- Apply single, fully specified multi-gene predictor of outcome to samples and categorize each patient as good or poor prognosis
- Categorizing each patient with regard to practice standards as requiring or not requiring chemotherapy
- Randomizing patients for whom genomic classifier and practice standards disagree with regard to use of chemotherapy
- Compare long term outcomes for randomized patients

Validation Study for Use of Chemotherapy in Node Negative Breast Cancer

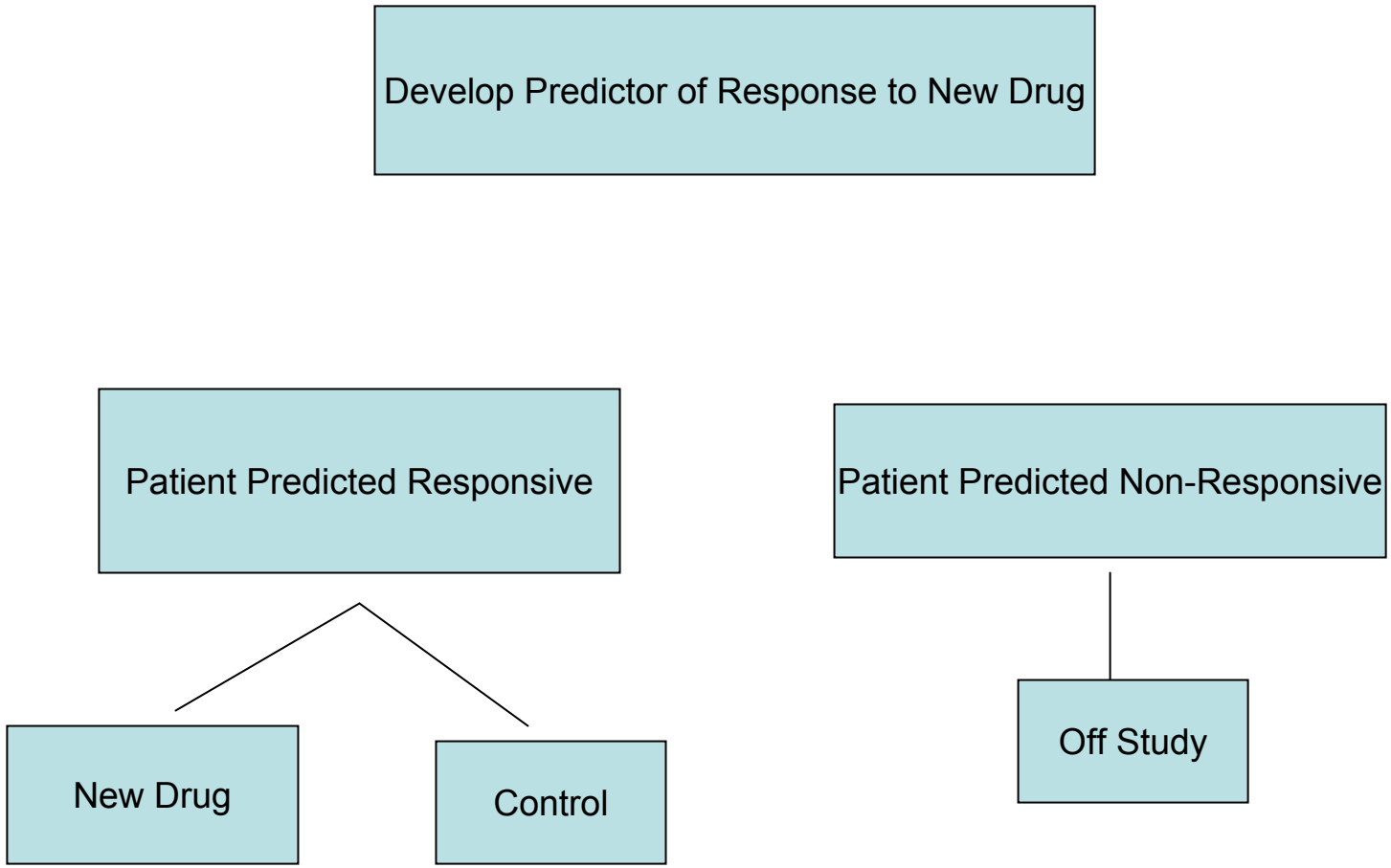
- Prospective study design
- Samples collected and assayed from patients with node negative ER+ breast cancer receiving TAM
- Identify patients predicted to be very good prognosis on TAM alone using a single, fully specified multi-gene predictor of outcome
- Were long-term outcomes for patients in good prognosis group sufficiently good to have warranted withholding chemotherapy?

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

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

Developmental Strategy (I)

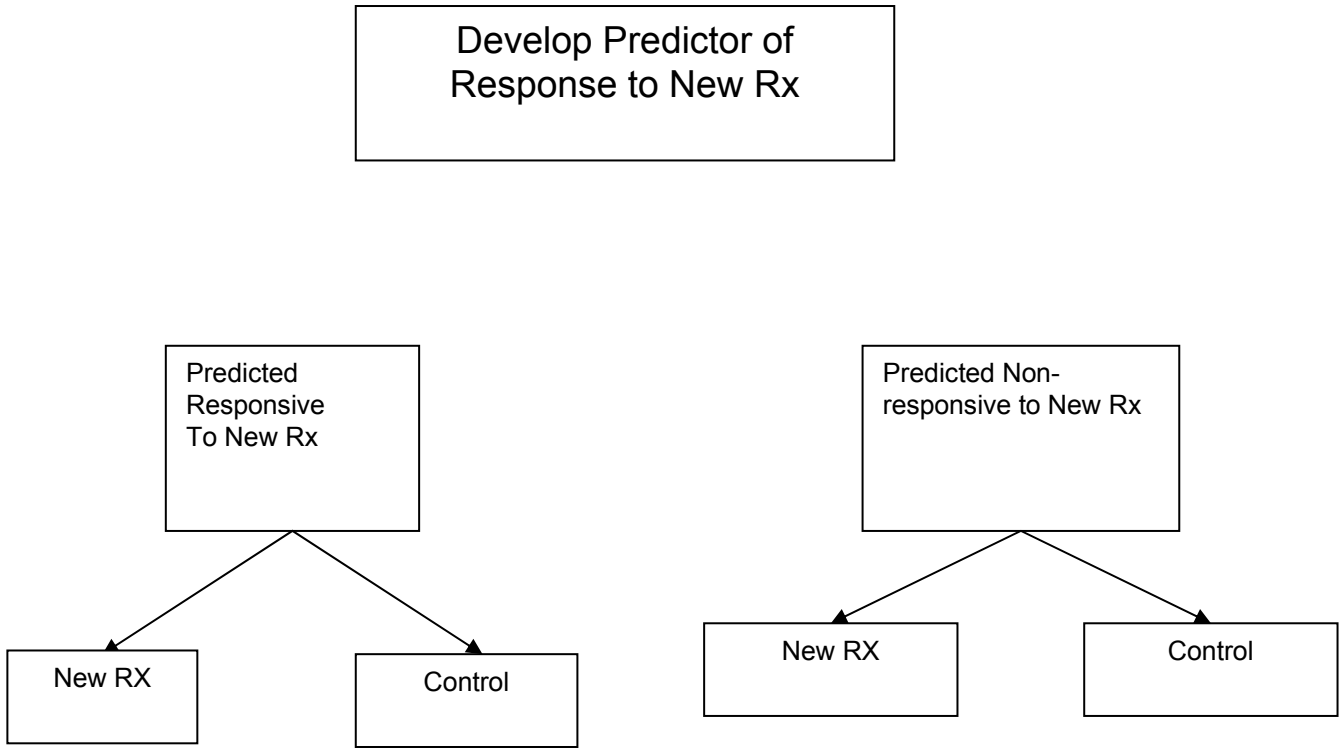
- **Develop** a diagnostic classifier that identifies the patients likely to benefit from the new drug
- Develop a reproducible assay for the classifier
- **Use** the diagnostic to restrict eligibility to a prospectively planned evaluation of the new drug
- Demonstrate that the new drug is effective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


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

Evaluating the Efficiency of Strategy (I)

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

Developmental Strategy (II)



Developmental Strategy (II)

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

Key Features of Design (II)

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

One Should Require That

- The classifier be reproducibly measurable
- The classifier in conjunction with the drug has clinical utility

There Should Be No Requirement For

- Demonstrating that the classifier or any of its components are FDA defined “validated biomarkers of disease status”
- Ensuring that the individual components of the classifier are correlated with patient outcome or effective for selecting patients for treatment
- Demonstrating that repeating the classifier development process on independent data results in the same classifier

The Roadmap

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

Guiding Principle

- The data used to develop the classifier must be distinct from the data used to test hypotheses about treatment effect in subsets determined by the classifier
  - Developmental studies are exploratory
  - 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