# Myths about the Development and Validation of Predictive Classifiers using Gene Expression Profiles

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

# “Biomarkers”

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

# Surrogate Endpoints

- It is extremely difficult to properly validate a biomarker as a surrogate for clinical outcome for use in phase III trials. It requires a series of randomized trials with both the candidate biomarker and clinical outcome measured
- Biomarkers for use in phase I/II studies need not be validated as surrogates for clinical outcome

# Predictive Biomarkers

- Most cancer treatments benefit only a minority of patients to whom they are administered
- 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

# Oncology Needs Predictive Markers not Prognostic Factors

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

- Criteria for validation of surrogate endpoints should not be applied to predictive biomarkers used for treatment selection

# Good Microarray Studies Have Clear Objectives

- Class Comparison
  - Find genes whose expression differs among predetermined classes, e.g. tissue or experimental condition
- Class Prediction
  - Prediction of predetermined class (e.g. treatment outcome) using information from gene expression profile
- Class Discovery
  - Discover clusters of specimens having similar expression profiles
  - Discover clusters of genes having similar expression profiles

# Class Comparison and Class Prediction

- Not clustering problems
- Supervised methods

# Class Prediction

- 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

# Components of Class Prediction

- Feature (gene) selection
  - Which genes will be included in the model
- Select model type
  - E.g. Diagonal linear discriminant analysis, Nearest-Neighbor, …
- Fitting parameters (regression coefficients) for model
  - Selecting value of tuning parameters

# Class Prediction ≠ Class Comparison

- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy.
- Statisticians are used to inference, not prediction
- Most statistical methods were not developed for p>>n prediction problems

# Myth

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

# Simple Gene Selection

- Use genes which are univariately correlated with outcome
  - For class comparison false discovery rate is important
  - For class prediction, predictive accuracy is important

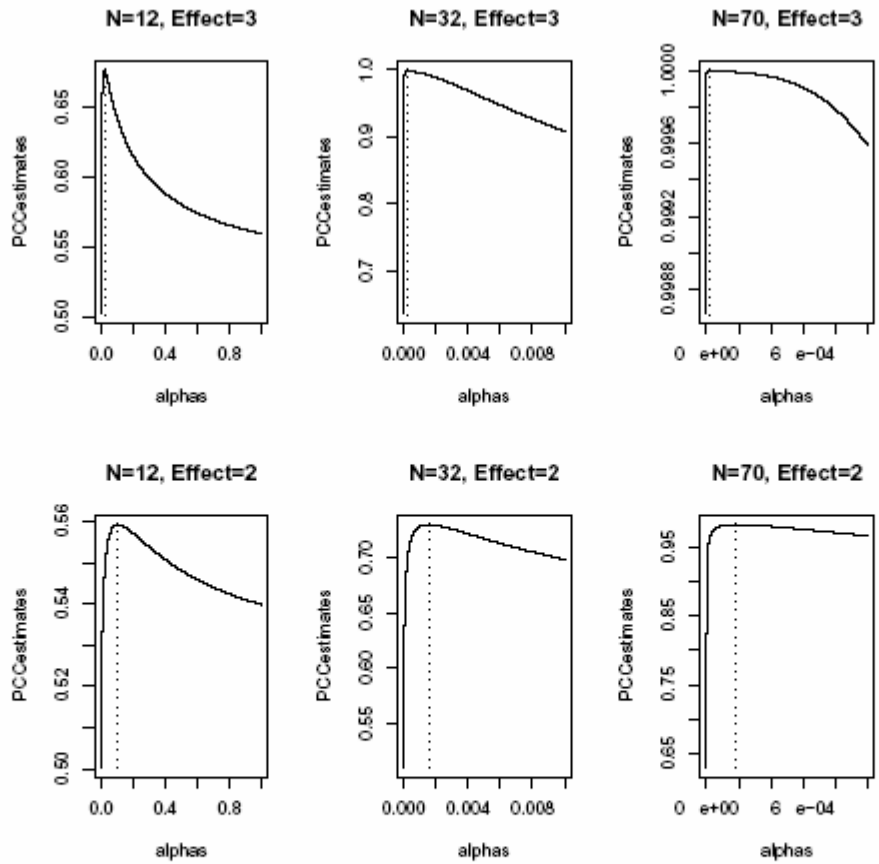



Figure 1: Plots of the estimated PCC as a function of $\alpha$, plotted for various values of $n$, based on Equation 10. In each plot, "Effect" is defined as $\frac{2\delta}{\sigma}$, $m = 10$ is the number of differentially

# Complex Gene Selection

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

# 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

# Linear Classifiers for Two Classes

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

# Other Simple Methods

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

# When p>>n

- It is always possible to find a set of features and a weight vector for which the classification error on the training set is zero.
- Why consider more complex models?

- Artificial intelligence sells to journal reviewers and peers who cannot distinguish hype from substance when it comes to microarray data analysis.
- Comparative studies generally indicate that simpler methods work as well or better for microarray problems because they avoid overfitting the data.

# Evaluating a Classifier

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

# Split-Sample Evaluation

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

# 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

# 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

- Cross validation is only valid if the test set is not used in any way in the development of the model. Using the complete set of samples to select genes violates this assumption and invalidates cross-validation.
- With proper cross-validation, the model must be developed *from scratch* for each leave-one-out training set. This means that feature selection must be repeated for each leave-one-out training set.
  - Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data. Journal of the National Cancer Institute 95:14-18, 2003.

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

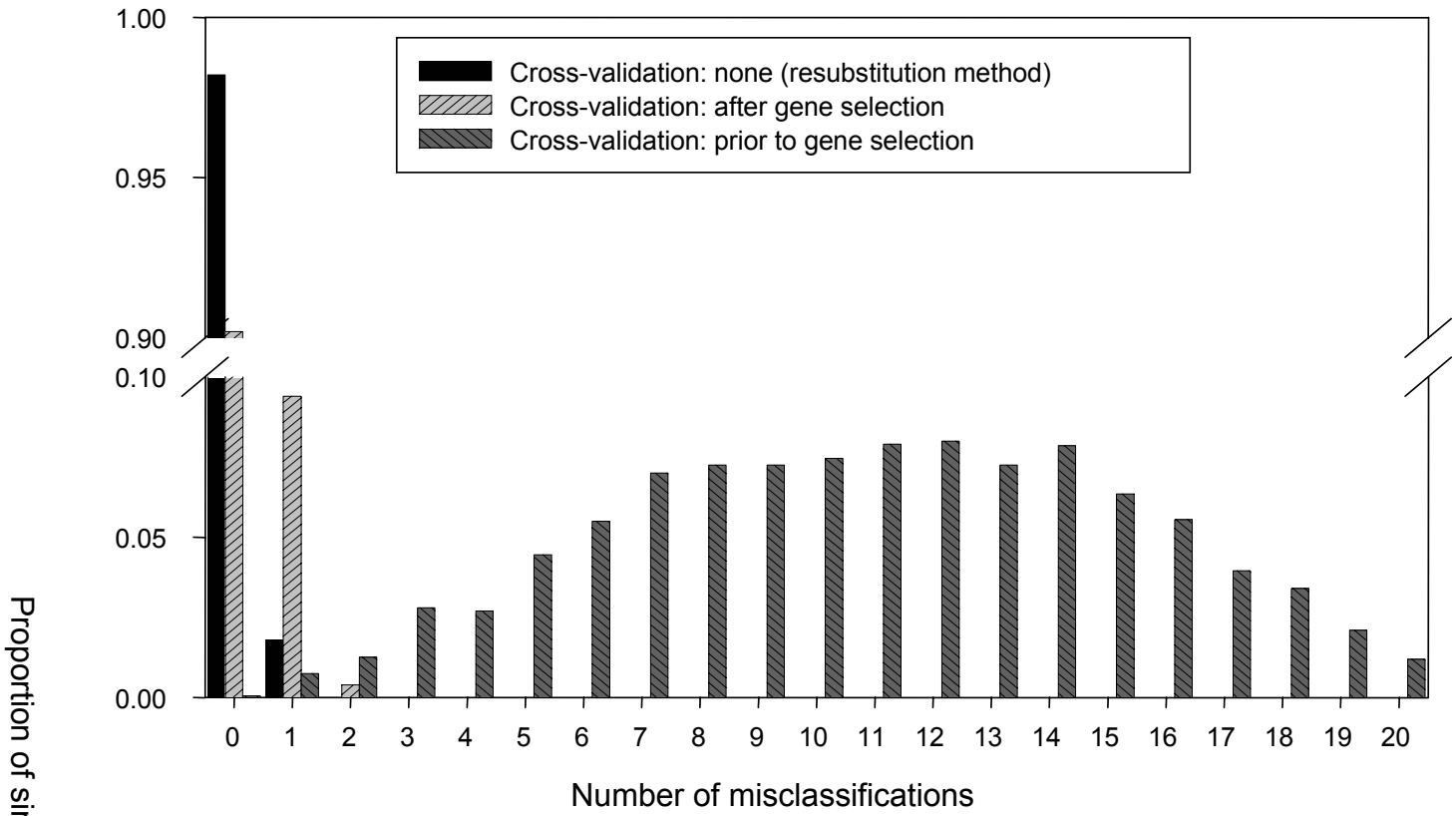
# Prediction on Simulated Null Data

**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 (*discussed later*)
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.


Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
1.00
0.95
0.90
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of misclassifications
Proportion of simulated data sets

# 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

# Simulated Data

## 40 cases, 10 genes selected from 5000

| Method | Estimate | Std Deviation |
|---|---|---|
| True | .078 | |
| Resubstitution | .007 | .016 |
| LOOCV | .092 | .115 |
| 10-fold CV | .118 | .120 |
| 5-fold CV | .161 | .127 |
| Split sample 1-1 | .345 | .185 |
| Split sample 2-1 | .205 | .184 |
| .632+ bootstrap | .274 | .084 |

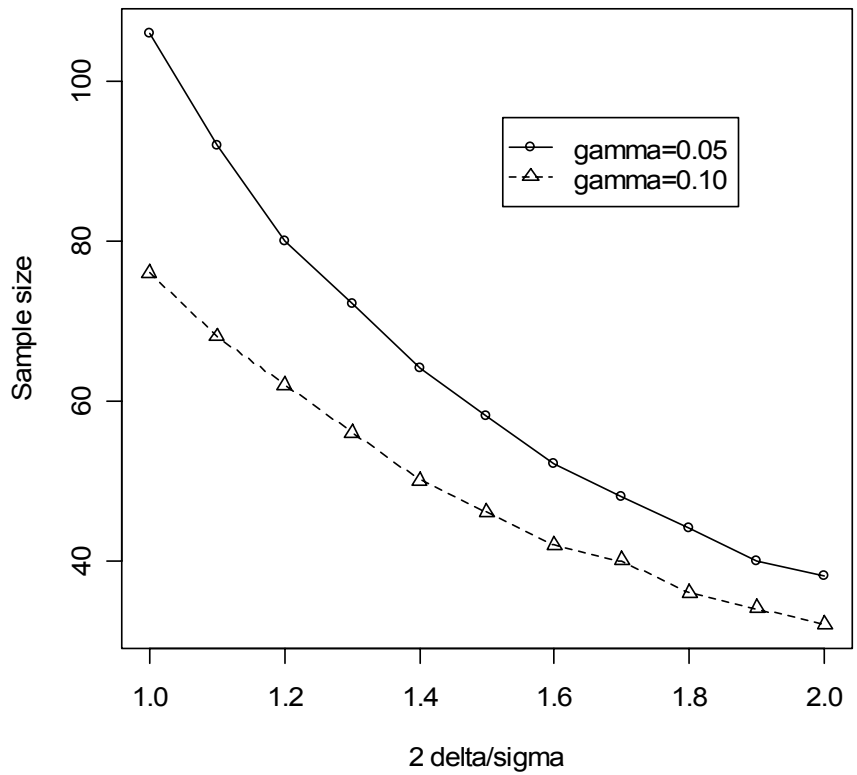
# Myth

- Huge sample sizes are needed to develop effective predictive classifiers

# Sample Size Planning References


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

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



# Myth

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

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

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

Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data. Journal of the National Cancer Institute 95:14-18, 2003.

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

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

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

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

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

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

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

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

Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research 10:6759-63, 2004.

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

Simon R. When is a genomic classifier ready for prime time? Nature Clinical Practice – Oncology 1:4-5, 2004.

Simon R. An agenda for Clinical Trials: clinical trials in the genomic era. Clinical Trials 1:468-470, 2004.

Simon R. Development and Validation of Therapeutically Relevant Multi-gene Biomarker Classifiers. Journal of the National Cancer Institute 97:866-867, 2005.

Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology 23(29), 2005.

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

Simon R. Guidelines for the design of clinical studies for development and validation of therapeutically relevant biomarkers and biomarker classification systems. In Biomarkers in Breast Cancer, Hayes DF and Gasparini G, Humana Press, pp 3-15, 2005.

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