Development & Validation of Genomic Classifiers for Treatment Selection

Richard Simon, D.Sc.

National Cancer Institute

http://linus.nci.nih.gov/brb

http://linus.nci.nih.gov/brb

- http://linus.nci.nih.gov/brb
  - Powerpoint presentations
  - Reprints & Technical Reports
  - BRB-ArrayTools software
  - BRB-ArrayTools Data Archive
  - Sample Size Planning for Targeted Clinical Trials

Good Microarray Studies Have Clear Objectives

- Class Comparison
  - Find genes whose expression differs among predetermined classes
- Class Prediction
  - Prediction of predetermined class (phenotype) 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

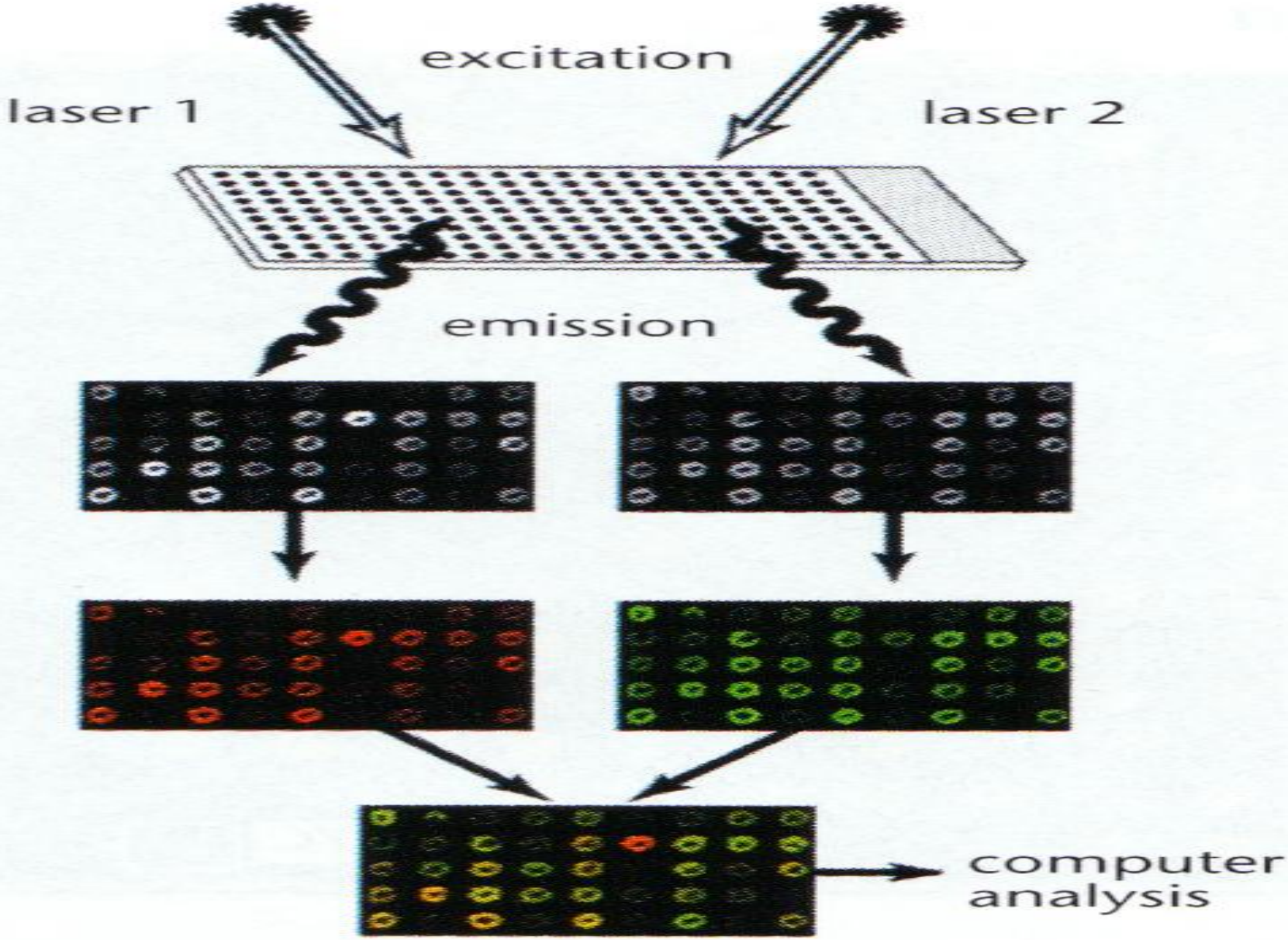
Class Prediction

- Predict which tumors will respond to a particular treatment
- Predict which patients will relapse after a particular treatment

Microarray Platforms for Developing Predictive Classifiers

- Single label arrays
  - Affymetrix GeneChips

- Dual label arrays using common reference design


excitation
laser 1
laser 2
emission
computer
analysis

- X_{ik}=expression of gene i in specimen from case k

- For single label arrays, expression is based on fluorescence intensity of gene i in specimen from case k
- For dual-label arrays, expression is based on log of ratio of fluorescence intensities of gene i in specimen from case k to that for common reference specimen

Class Prediction Model

- Given a sample with an expression profile vector *x* of log-ratios or log signals and unknown class.
- Predict which class the sample belongs to
- The class prediction model is a function *f* which maps from the set of vectors *x* to the set of class labels {1,2} (if there are two classes).
- *f* generally utilizes only some of the components of *x* (i.e. only some of the genes)
- Specifying the model *f* involves specifying some parameters (e.g. regression coefficients) by fitting the model to the data (*learning* the data).

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

Gene Selection

- Genes that are differentially expressed among the classes at a significance level α (e.g. 0.01)
  - The α level is selected only to control the number of genes in the model
    - For class comparison false discovery rate is important
    - For class prediction, predictive accuracy is important

Estimation of Within-Class Variance

- Estimate separately for each gene σ_g^2
- Assume all genes have same variance
- Random (hierarchical) variance model
  - Wright G.W. and Simon R. Bioinformatics19:2448-2455,2003
  - Inverse gamma distribution of residual variances
  - Results in exact F (or t) distribution of test statistics with increased degrees of freedom for error variance
  - For any normal linear model

Gene Selection

- Small subset of genes which together give most accurate predictions
  - Combinatorial optimization algorithms
    - 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

$$\underline{w} = y' S^{-1}$$

  - Requires estimating correlations among all genes selected for model
  - y = vector of class mean differences
- Diagonal linear discriminant analysis (DLDA) assumes features are uncorrelated
- Compound covariate predictor (Radmacher) and Golub's method are similar to DLDA

Linear Classifiers for Two Classes

- Support vector machines with inner product kernel are linear classifiers with weights determined to separate the classes with a hyperplain that minimizes the length of the weight vector

Support Vector Machine

minimize $\sum_{i} w_i^2$

subject to $y_j\left(\underline{w}'\underline{x}^{(j)} + b\right) \geq 1$

where $y_j = \pm 1$ for class 1 or 2.

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?

Myth

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

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

Other Simple Methods

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

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

Evaluating a Classifier

- The classification algorithm includes the following parts:
  - Determining what type of classifier to use
  - Gene selection
  - Fitting parameters
  - Optimizing with regard to tuning parameters
- If a re-sampling method such as cross-validation is to be used to estimate predictive error of a classifier, **all** aspects of the classification algorithm must be repeated for each training set and the accuracy of the resulting classifier scored on the corresponding validation set

- 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

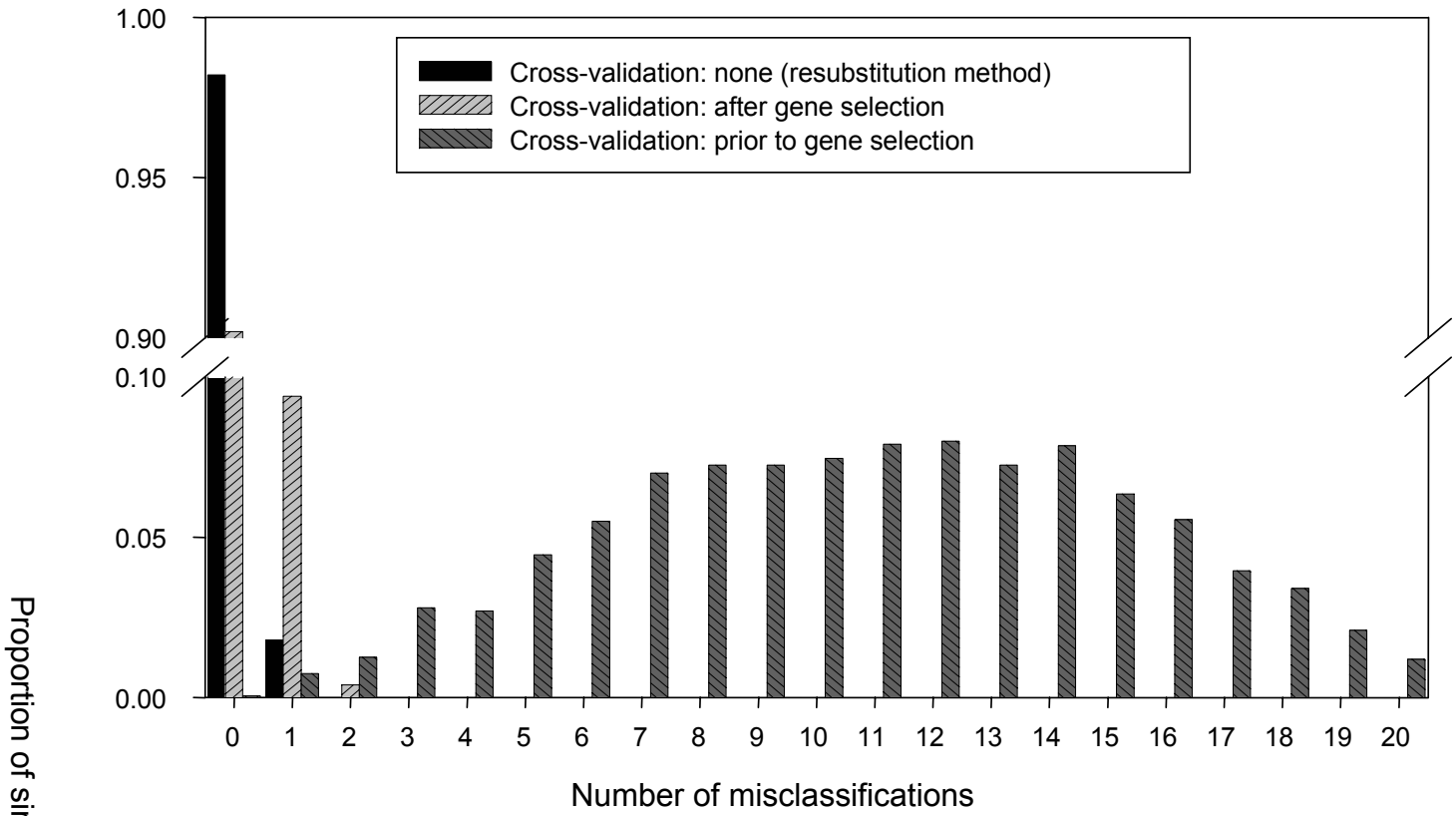
Prediction on Simulated Null Data

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

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

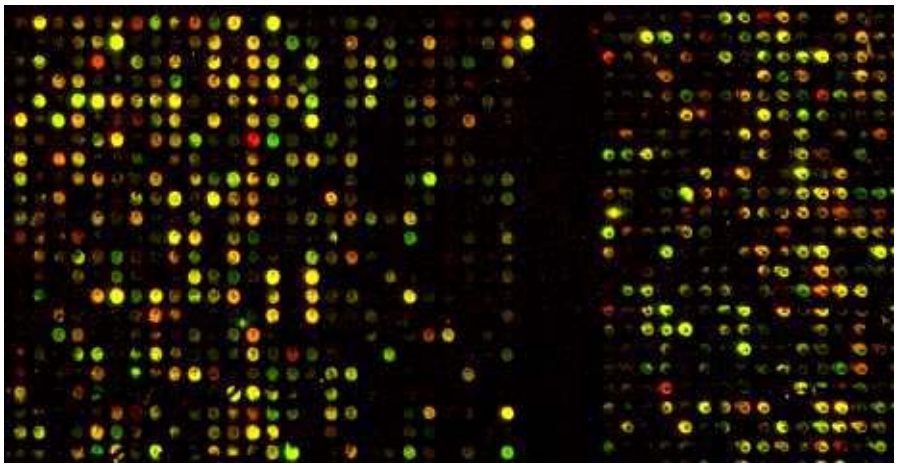
Permutation Distribution of Cross-validated Misclassification Rate of a Multivariate Classifier

- 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

Gene-Expression Profiles in Hereditary Breast Cancer

cDNA Microarrays

Parallel Gene Expression Analysis



- Breast tumors studied:
  - 7 *BRCA1+* tumors
  - 8 *BRCA2+* tumors
  - 7 sporadic tumors

- Log-ratios measurements of 3226 genes for each tumor after initial data filtering

RESEARCH QUESTION

Can we distinguish *BRCA1+* from *BRCA1–* cancers and *BRCA2+* from *BRCA2–* cancers based solely on their gene expression profiles?

Classification of BRCA2 Germline Mutations

Classification Method	LOOCV Prediction Error
Compound Covariate Predictor	14%
Fisher LDA	36%
Diagonal LDA	14%
1-Nearest Neighbor	9%
3-Nearest Neighbor	23%
Support Vector Machine (linear kernel)	18%
Classification Tree	45%

Common Problems With Cross Validation

- Pre-selection of genes using entire dataset
- Failure to consider optimization of tuning parameter part of classification algorithm
  - Varma & Simon, BMC Bioinformatics 7:91 2006

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

Survival Risk Group Prediction

- Define algorithm for selecting genes and constructing survival risk groups
- Apply algorithm in LOOCV fashion to obtain predicted survival risk groups
- Compute Kaplan-Meier curves for cross-validated risk groups
- Compute permutation p value for separation of cross-validated Kaplan-Meier curves
- Compare separation of cross-validated Kaplan-Meier curves to separtion of K-M curves for standard clinical staging
- Available in BRB-ArrayTools
  - http://linus.nci.nih.gov/brb

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)

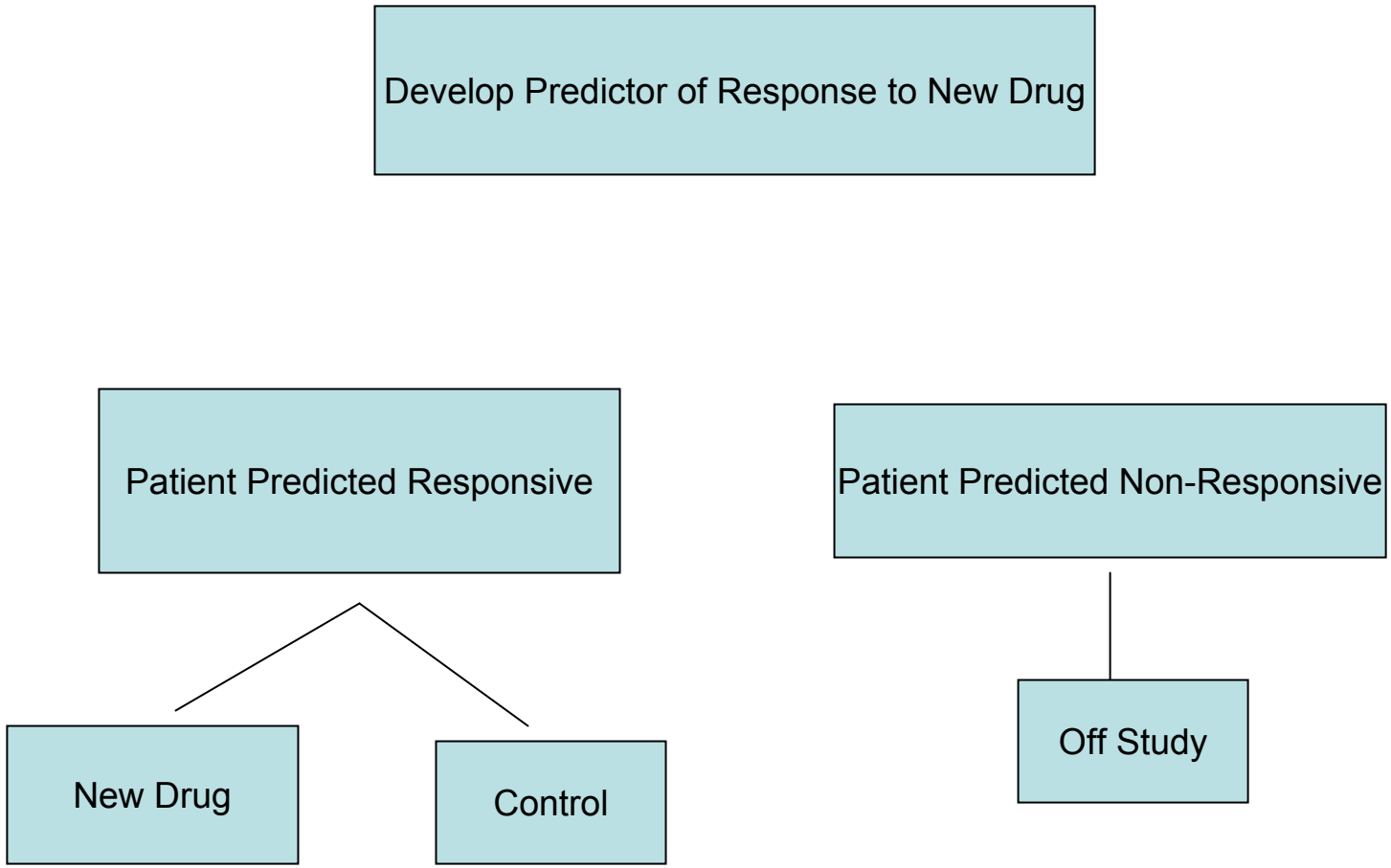
External Validation

- Should address clinical utility, not just predictive accuracy
- Should incorporate all sources of variability likely to be seen in broad clinical application

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

Developmental Strategy

- **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

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

Acknowledgements

- Kevin Dobbin
- Michael Radmacher
- Sudhir Varma
- Annette Molinaro

Selected Features of BRB-ArrayTools linus.nci.nih.gov/brb

- Multivariate permutation tests for class comparison to control number and proportion of false discoveries with specified confidence level
- Fast implementation of SAM
- Extensive annotation for genes
- Find genes correlated with censored survival while controlling number or proportion of false discoveries
- Gene set comparison analysis
- Analysis of variance (fixed and mixed)

Selected Features of BRB-ArrayTools

- Class prediction
  - DLDA, CCP, Nearest Neighbor, Nearest Centroid, Shrunken Centroids, SVM, Random Forests,Top scoring pairs
  - Complete LOOCV, k-fold CV, repeated k-fold, .632+ bootstrap
  - permutation significance of cross-validated error rate
- Survival risk group prediction
- R plug-ins