Developing and Evaluating Medical Diagnostics Based on Whole Genome Technologies

Richard Simon

National Cancer Institute

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

- Clinical trial for patients with breast cancer, without nodal or distant metastases, Estrogen receptor positive tumor
  - 5 year survival rate for control group (surgery + radiation + Tamoxifen) expected to be 90%
  - Size trial to detect 92% survival in group treated with control modalities plus chemotherapy

Treating the Many for the Benefit of the Few

- Acceptable to many physicians, companies and statisticians
  - Broad eligibility
  - Avoid subset analysis
- Not so good for patients or for their health budget

Using DNA Microarrays to Select Patients for Phase III Trials

- Perform microarray gene expression profiling on patients in phase II trials of new drug E
- Develop gene expression based predictor of responsiveness to E
- Select patients for phase III trial based on predicted responsiveness to E

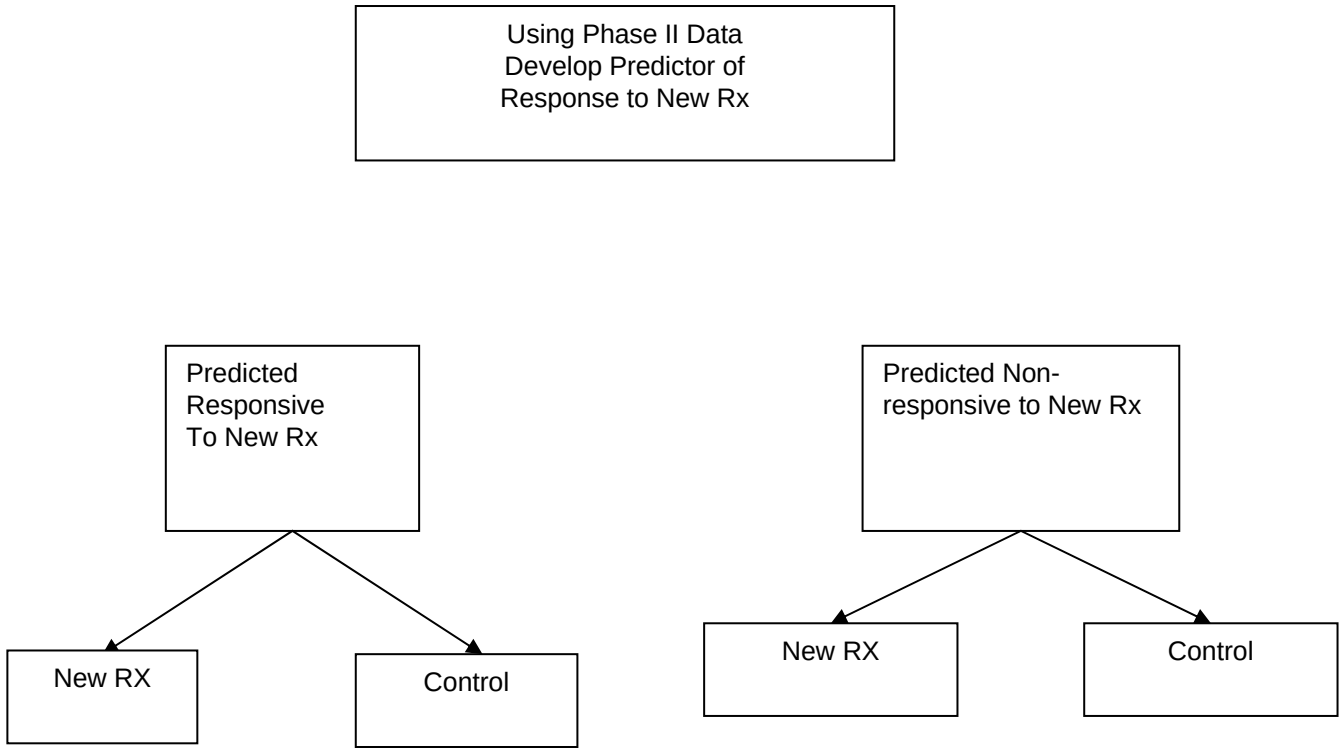
Randomized Clinical Trials Targeted to Patients Predicted to be Responsive to the New Treatment Can Be Much More Efficient than Traditional Untargeted Designs

- Simon R and Maitnourim A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research (In Press)
- Maitnourim A and Simon R. On the efficiency of targeted clinical trials. Statistics in Medicine (In Press).

- For a drug like Iressa in lung cancer
  - 10% response rate
  - Traditional untargeted designs are very inefficient, even with 1000 patients randomized
  - More effort should be placed in finding predictors of response based on phase II data
    - Sequencing key genes
    - Expression profiling

- For Herceptin, even a relatively poor assay enabled conduct of a targeted phase III trial which was crucial for establishing effectiveness
- In many cases, the assay based on the presumed mechanism of action will not correlate with response and it may be more effective to let the data develop the assay via expression profiling

Using DNA Microarrays to Select Patients for Phase III Trials



Studies of Prognostic Classifiers

- Contradictory literature
- Few prognostic classifiers utilized in therapeutic decision making

Common Problems With Developmental Studies of Diagnostic Classifiers

- Convenience sample of available specimens
- No prospectively stated hypotheses, protocol for patient selection or analysis plan
- Results are often not medically relevant because of patient heterogeneity
    - Eg mixture of N+, N-, with & without chemotherapy
- Multiple comparisons without structure or statistical control leading to non-reproducibility of findings
- Confounding of tissue handling and assay procedures with outcomes

Traditional Problems + New Statistical Problems in Dealing With High Dimensional Data

Using Microarray Expression Profiles for Class Prediction

- Predict membership of a specimen in pre-defined classes
  - Responders vs non-responders to a treatment
  - Toxic reaction vs no-toxic reaction

Most Statistical Methods Are For Inference, Not Prediction and Particularly Not for p>>n Prediction Problems

- p = number of *candidate* predictors
- Development and validation of diagnostic classifiers are primarily problems of prediction, not of inference about parameters
  - Predictive accuracy, not false positive genes, statistical significance or goodness of fit
- Demonstrating predictive accuracy on the data used for model development is not adequate
  - With p>>n, re-substitution error of zero is always possible

Components of Class Prediction Algorithm

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

Feature 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; the number of false positives is not directly relevant
  - Class prediction is different than class comparison

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)
  - Ignores correlations among genes
- Compound covariate predictor
- Golub's weighted voting method
- Partial least squares
- Support vector machines with inner product kernel
- Perceptrons (neural networks with no hidden layer)

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.

The Set of Linear Models is Too Rich for p>>n Classification

- 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
  - All p>>n problems are linearly separable
- How to select a linear classifier
- 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.

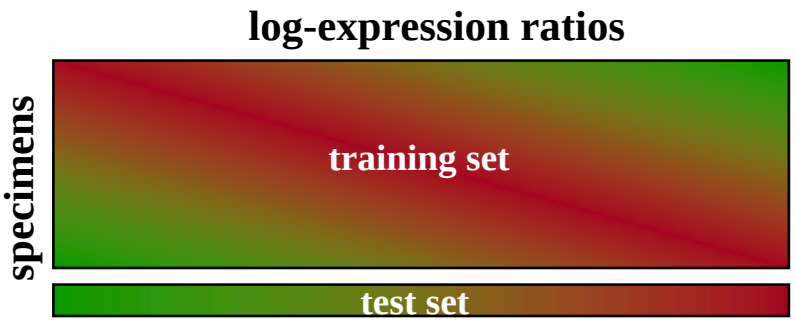
Other Simple (and effective) Methods

- Nearest neighbor classification
- Nearest centroid classification
- Shrunken centroid classification

Internal Validation of a Classifier

- Re-substitution estimate
  - Develop classifier on dataset, test predictions on same data
  - Very biased for p>>n
- Split-sample validation
  - Split data into training and validation sets
  - Test *single fully specified model* on the validation set
  - Often applied with too small a validation set
  - Often applied invalidly with tuning parameter optimized on validation set
- Cross-validation

Cross-Validated Prediction (Leave-One-Out Method)



1. Full data set is divided into training and validation sets (validation set contains 1 specimen).
2. Prediction rule is built *from scratch* using the training set.
3. Rule is applied to the specimen in the validation set.
4. Process is repeated until each specimen has appeared once in a validation set.

- Cross validation is only valid if the test validation 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

- For small studies, cross-validation, if performed correctly, can be preferable to split-sample validation
  - Cross-validation can only be used when there is a well specified algorithm for classifier development
- Internal validation is limited by
  - Limited precision of estimated error rate
  - Limitations of data used for developmental study

Common Limitations in Data Used for Internal Validation

- Heterogeneity of patients
  - Associations not therapeutically relevant
  - Associations due to un-modeled variables
- Confounding of profiles with sample handling or assay parameters, including assay drift
- Failure to reflect sources of assay variability that will exist in broad clinical application

External Validation

- Specimens from prospective multi-center clinical trial
- Specimens assayed at different time from training data
- Positive and negative samples handled in the same way and assayed blinded to outcome
- Study sufficiently large to give reasonable precise estimate of sensitivity and specificity of the multivariate classifier
- The validation study is prospectively planned
  - patient selection pre-specified to address a therapeutically relevant question
  - endpoints and hypotheses pre-specified
  - predictor fully pre-specified
  - Study addresses assay reproducibility
  - Specimens may be either prospective or archived

Steps in Development of Therapeutically Relevant Genomic Diagnostics

- Phase I: Unstructured study to show that biomarkers have relevance to the disease
- Phase II:
  - Select therapeutically relevant population
    - Node negative, ER+, well staged breast cancer patients who have received Tamoxifen alone
  - Perform genome wide expression profiling of patients in large clinical trials using frozen archived material to develop profile classifier of outcome or treatment benefit
    - Obtain unbiased internal estimate of prediction accuracy
- Adapt platform for broad clinical application
- Establish assay reproducibility
- Phase III: External validation of fully specified profile classifier in prospectively planned analysis
  - of previously performed clinical trial using archived blocks
  - of new clinical trial in which the classifier is used in real time