Validation of Therapeutically Relevant Diagnostic Classifiers Based on Gene Expression Data

Richard Simon

National Cancer Institute

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

Resources

- Biometric Research Branch Website http://linus.nci.nih.gov/brb
  - Powerpoint presentation
  - Reprints & Technical Reports
  - BRB-ArrayTools software
- *Design and Analysis of DNA Microarray Investigations*
  - R Simon, EL Korn, MD Radmacher, L McShane, G Wright, Y Zhao. Springer (2003)

Why We Need Diagnostic Classifiers

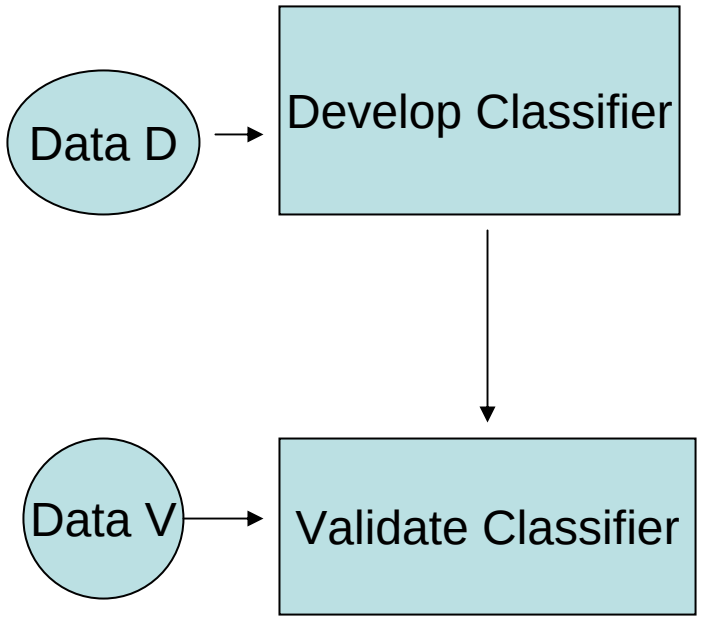
- Clinical trial for patients with breast cancer, without nodal or distant metastases, Estrogen receptor positive tumor
  - 5 year survival rate for control group (surgery + radiation + Tamoxifen) expected to be 90%
  - Size trial to detect 92% survival in group treated with control modalities plus chemotherapy
  - Only 2% of patients benefit

Common Problems With Diagnostic Classifiers

- Not therapeutically relevant
- Not reliable
- Not validated

Many Microarray Studies Do Not Address A Medically Relevant Question

- Comparing expression in AML vs ALL
- Finding genes whose expression correlates with outcome in a heterogeneous group of primary breast cancer patients is usually not therapeutically meaningful
  - N+, N-, ER+, ER-, +- chemotherapy


Data D
Develop Classifier
Data V
Validate Classifier

What is a Classifier?

- Given a vector of "features" x_1, x_2, …,x_p for a case to be classified
- The classifier provides a prediction of the "class" that the case belongs to
- The classifier is designed to be used for decision making in a specific medical context
- E.g. 2 classes, responders & non-responders to a specified treatment

Components of Classifier Development

- Select cases and classes for training data
- Feature (gene) selection
  - Which genes will be included in the classifier
- Select model type
  - E.g. Diagonal linear discriminant analysis, Nearest-Neighbor, Neural network, …
- Fitting parameters of model to data from training set (*training*)

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

Classifier Development

- Will not address classifier development here except to caution that:
- The literature on classifier development is dense with hype
- The classes in all training sets with p > n are perfectly linear separable
  - p= number of candidate genes
  - n= number of cases
- There is rarely enough information available for utilizing non-linear classifiers without over-fitting 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.
- When the number of candidate predictors (p) exceeds the number of cases (n), perfect prediction on the same data used to create the predictor is always possible

- The vast literature of studies developing prognostic or predictive markers “validate” their models on the same set of data used to develop the models.
- That is one reason why that literature is so non-reproducible

Validation of *Biomarker* Classifiers

- We only care about biomarker classifiers if they help us care for patients
- It is a serious error to try to validate a biomarker in an absolute sense rather than evaluating its use for therapeutic decision making in a specific medical context

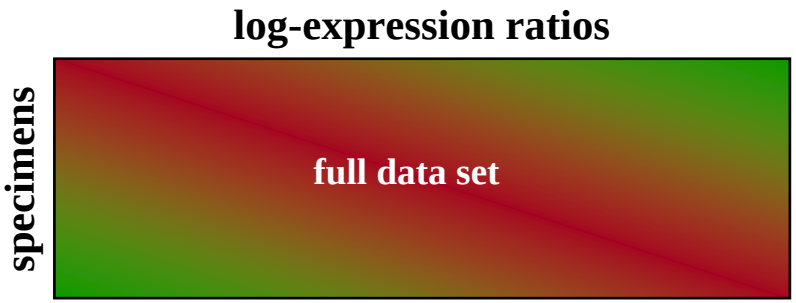
Internal Validation of a Classifier

- Resubstitution estimate
  - Develop classifier on dataset, test predictions on same data
  - Horribly biased for p>>n
- Split-sample validation
  - Split data into training and test sets
  - Test *single fully specified model* on the test set
- Cross-validation

Split-Sample Evaluation

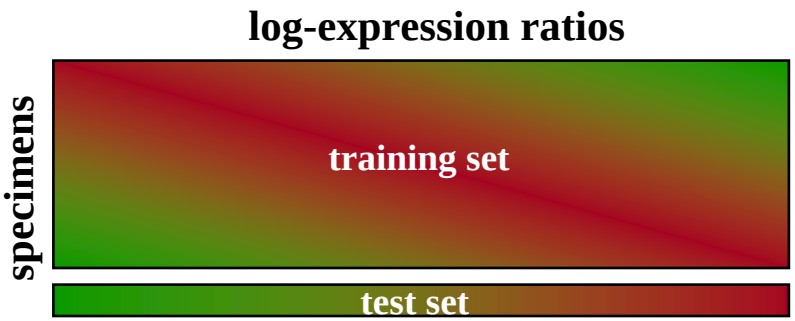
- Used for Rosenwald et al. study of prognosis in DLBL lymphoma.
  - 200 cases training-set
  - 100 cases test-set

Non-Cross-Validated Prediction



1. Prediction rule is built using full data set.
2. Rule is applied to each specimen for class prediction.

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



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

- 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, complete cross-validation gives more precise estimates of prediction error than with split-sample validation
  - Cross-validation can only be used when there is a well specified algorithm for classifier development
- Internal validation is limited by
  - Sample size of the data available
  - Lack of representation of important real world sources of variability in data used for developmental study

Common Limitations in Data Used for Internal Validation

- Confounding by sample handling or assay effects
  - Cases collected and assayed at different times than controls
- Failure to incorporate important sources of future variability
  - Assay drift
  - Inter-lab variability
- Change in distribution of un-modeled variables

External Validation

- Specimens from prospective multi-center clinical trial
- Specimens assayed at different time from training data in a manner that simulates mode of subsequent use
- 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

Prospective/Retrospective Validation Study Node negative Breast Cancer GHI / NSABP

- Prospective study design
- Samples collected and archived from patients with node negative ER+ breast cancer receiving TAM (B14)
- Apply single, fully specified multi-gene RT-PCR classifier of outcome to samples and predict good or poor outcome for each patient
  - Classifier developed previously using microarray analysis of other data and transferred to RT-PCR platform
- Are long-term outcomes for patients in good prognosis group sufficiently good to withhold chemotherapy?

Prospective Validation Design

- Randomize patients with node negative ER+ breast cancer receiving TAM to chemotherapy vs classifier determined therapy
- Determine whether classifier determined arm has equivalent outcome to arm in which all patients receive chemotherapy
  - Therapeutic equivalence trial
- Gold standard but rarely performed
  - Very inefficient because most patients get same treatment in both arms and so the trial must be sized to detect miniscule difference in outcome

Prospective Validation Design

- Randomize only patients with node negative ER+ breast cancer receiving TAM who are predicted to have good outcome using the classifier
  - Half randomized to chemotherapy, half to no chemotherapy
- Determine whether two arms have equivalent outcome
  - Therapeutic equivalence trial but more efficient because all patients in the two arms receive different treatment

Validation Study for Identifying Node Positive Patients Who Benefit from a Specific Regimen

- Standard treatment C
- New treatment E
- Predictor based on previous data for identifying patients who benefit from E but not C
- Randomized study of E vs C
- Measure gene expression on all patients
- Compare E vs C separately within groups predicted to benefit from E and those not predicted to benefit from E
- Two clinical trials worth of patients

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).

- Pre-prints available at http://linus.nci.nih.gov/brb

Steps in Development of Therapeutically Relevant Genomic Diagnostics

- Select therapeutically relevant population
  - Node negative, ER+, well staged breast cancer patients who have received Tam alone and have long follow-up
- 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
- 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