Strengths & Weaknesses of Observational Validation Designs for High Dimensional Data

Richard Simon, D.Sc.

National Cancer Institute

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

- Problematic to combine validation for diagnostic classification with validation for early detection
  - For early detection, use of biomarker introduces new subject management options
  - For some diagnostic classification contexts, therapeutic options are the same with or without use of biomarker

- Problematic to talk about biomarker validation
  - We only care about biomarkers if they help us care for patients or prevent people from becoming patients
  - It is an important error to try to validate a biomarker rather than evaluating use of a biomarker in a given manner in a specific medical context

- There is confusion on how to properly design and analyze studies involving high dimensional assays such as DNA microarrays
- This is compounds the serious traditional problems with the effective development and validation of diagnostic markers

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

- Development and validation of diagnostic classifiers are primarily problems of prediction, not of inference about parameters
- Demonstrating goodness of fit of a model to the data used to develop it is not a demonstration of predictive accuracy
  - With p>>n, perfect goodness of fit is always possible
- Many standard statistical methods are not effective for p>>n prediction problems
- *Design & Analysis of DNA Microarray Experiments*, Simon, Korn, McShane, Radmacher, Wright, Zhao

Traditional Approach for Marker Development

- Focus on candidate protein involved in disease pathogenesis
- Develop assay
- Conduct retrospective study of whether marker is prognostic using available specimens
- Marker dies because
  - Therapeutic relevance not established
  - Adequate validation study not performed
  - Inter-laboratory reproducibility not established
  - Contradictory reports on assay value

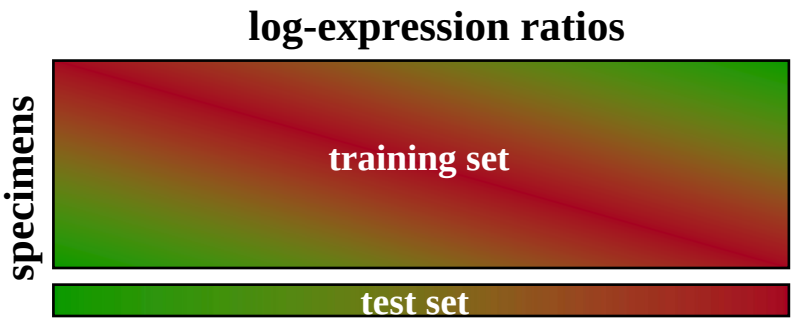
Common Problems With Developmental Studies of Diagnostic Classifiers

- 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 or case/control identifiers

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
  - Often applied with too small a validation set
- Cross-validation

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.

- 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

- For small studies, cross-validation, if performed correctly, is preferable to split-sample validation
  - Cross-validation can only be used when there is a well specified algorithm for classifier development
  - Journals are publishing studies based on split sample validation with meaninglessly small validation sets while rejecting studies based on cross-validation
- 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
  - Multiple analyses
- Confounding of profiles with sample handling or assay reagent effects, 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

- 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

Validation Study
Node negative Breast Cancer

- Prospective study design
- Samples collected and archived from patients with node negative ER+ breast cancer receiving TAM
- Apply single, fully specified multi-gene predictor of outcome to samples and categorize each patient as good or poor prognosis
- Are long-term outcomes for patients in good prognosis group sufficiently good to withhold chemotherapy?

Prospectively Planned Validation Using Archived Materials

- Fully specified classifier applied prospectively to frozen specimens from NSABP B14 patients who received Tamoxifen for 5 years
- Long term follow-up available
- Good risk patients had very good relapse-free survival

Prospective Validation Designs

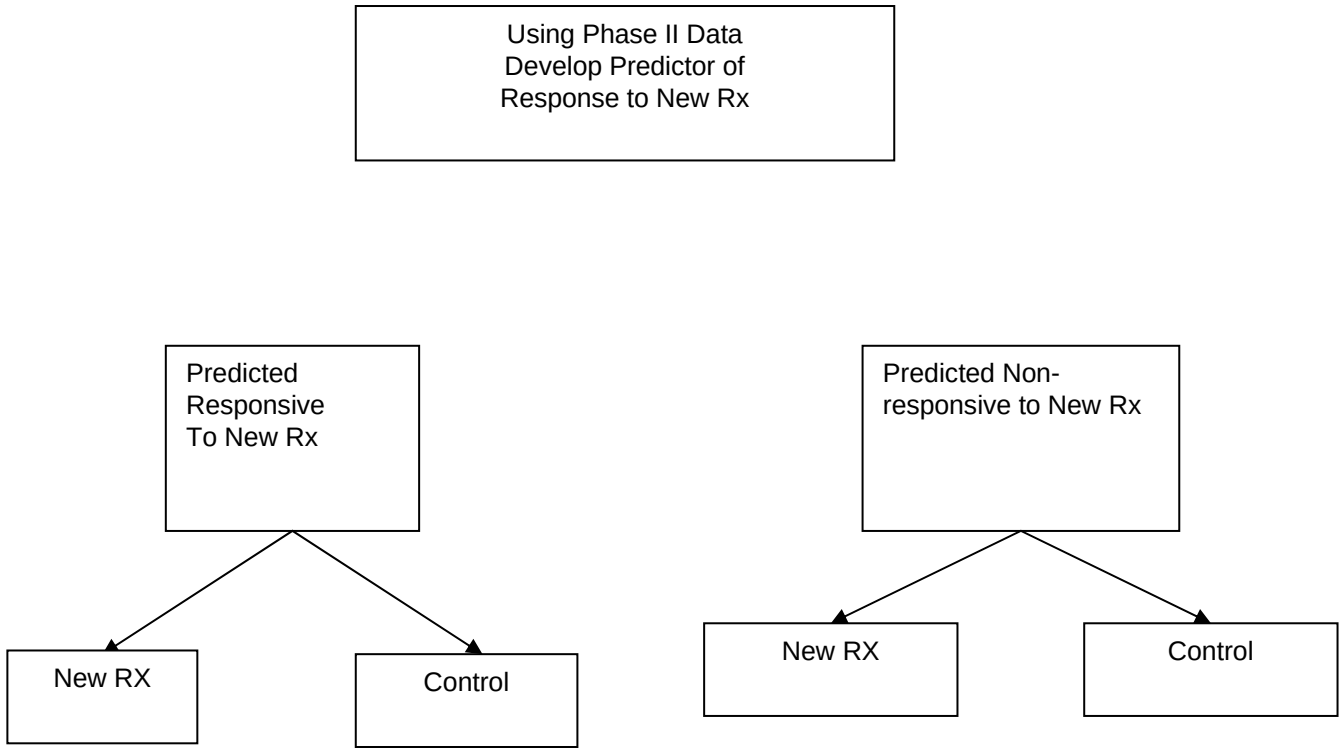
- Randomize patients to chemotherapy vs classifier determined therapy
- Gold standard but rarely performed
  - Very inefficient

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

Using DNA Microarrays to Select Patients for Phase III Trials



- For a drug like Iressa in lung cancer
  - 10% response rate
  - If only responders benefit, 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