A Roadmap for Developing & Validating Genomic Classifier for Treatment Selection

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

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), September 6, 2005.

Freidlin B and Simon R. Adaptive signature design. Clinical Cancer Research (In Press).

Simon R. Validation of pharmacogenomic biomarker classifiers for treatment selection. Disease Markers (In Press).

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 (In Press).

BRB-ArrayTools

- Integrated software package using Excel-based user interface but state-of-the art analysis methods programmed in R, Java & Fortran
- Publicly available for non-commercial uses from BRB website
- Licenseable from NIH Office of Technology Transfer for commercial uses

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

Why are most prognostic factors not used?

- Not therapeutically relevant
- Results do not appear reproducible

Why are most prognostic factors not therapeutically relevant?

- Patient selected for analysis are too heterogeneous; They do not represent a cohort for which a meaningful therapeutic question can be addressed

Why are most prognostic factor studies not reliable?

- They violate the key principle of separating model development from model evaluation
- Exploratory vs validation
- Issues of multiple testing and model over-fitting in exploratory studies
- Insufficient sample size and patient selection to develop therapeutically relevant model
- Do not adequately address assay reproducibility

- What is a genomic patient classification system?
- How does it differ from a set of prognostic genes?

A set of genes is not a classifier

- Gene selection
- Mathematical function for mapping from multivariate gene expression domain to prognostic or diagnostic classes
- Weights and other parameters including cut-off thresholds for risk scores

Linear Classifiers for Two Classes

$$l(\underline{x}) = \sum_{i \varepsilon G} w_i x_i$$

$\underline{x} =$ vector of expression measurements

$G =$ genes included in model

$w_i =$ weight for i'th gene

decision boundary $l(\underline{x})$ > or < d

There Should Be No Requirement For

- Demonstrating that the classifier or any of its components are “validated biomarkers of disease status”
- Demonstrating that repeating the classifier development process on independent data results in the same classifier

One Should Require That

- The classifier be reproducibly measurable

- The classifier in conjunction with the medical product has clinical utility

Design of Studies that Develop Therapeutically Relevant Genomic Classifiers

- How do you select appropriate patients?
- How many patients are needed?

Split Sample Approach

- Separate training set of patients from test set
- Patients should represent those eligible for a clinical trial that asks a therapeutically relevant question
- Do not access information about patients in test set until a single completely specified classifier is agreed upon based on the training set data

- If clinical trial is single arm test of new regimen and objective is to identify patients who are likely to respond:
  - A minimum of 15 responders and 15 non-responders should be included in the training set (Dobbin & Simon, in preparation)
  - Number of patients in test set should be determined to establish confidence limits for positive and negative predictive values of classifier consistent with clinical relevance

- If clinical trial is single arm study of standard therapy with time-to-event endpoint to identify patients at high risk of relapse, above considerations imply minimum of 15 patients with relapse.

- To identify patients who preferentially benefit from treatment A compared to treatment B with either response or time-to-event endpoints, a larger number of patients are needed for the training set.

Re-Sampling Approach

- Partition data into training set and test set
- Develop a single fully specified classifier of outcome on training set
- Use the classifier to predict outcome for patients in the test set and estimate the error rate
- Repeat the process for many random training-test partitions

- Re-sampling is only valid if the training 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 the process
- With proper re-sampling, the model must be developed from scratch for each training set. This means that gene selection must be repeated for each training set.

- Re-sampling, e.g. leave-one-out cross-validation is widely misunderstood even by statisticians and widely misused in the published clinical literature
- It is only applicable when there is a completely pre-defined algorithm for gene selection and classifier development that can be applied blindly to each training set

Developmental vs Confirmatory Studies

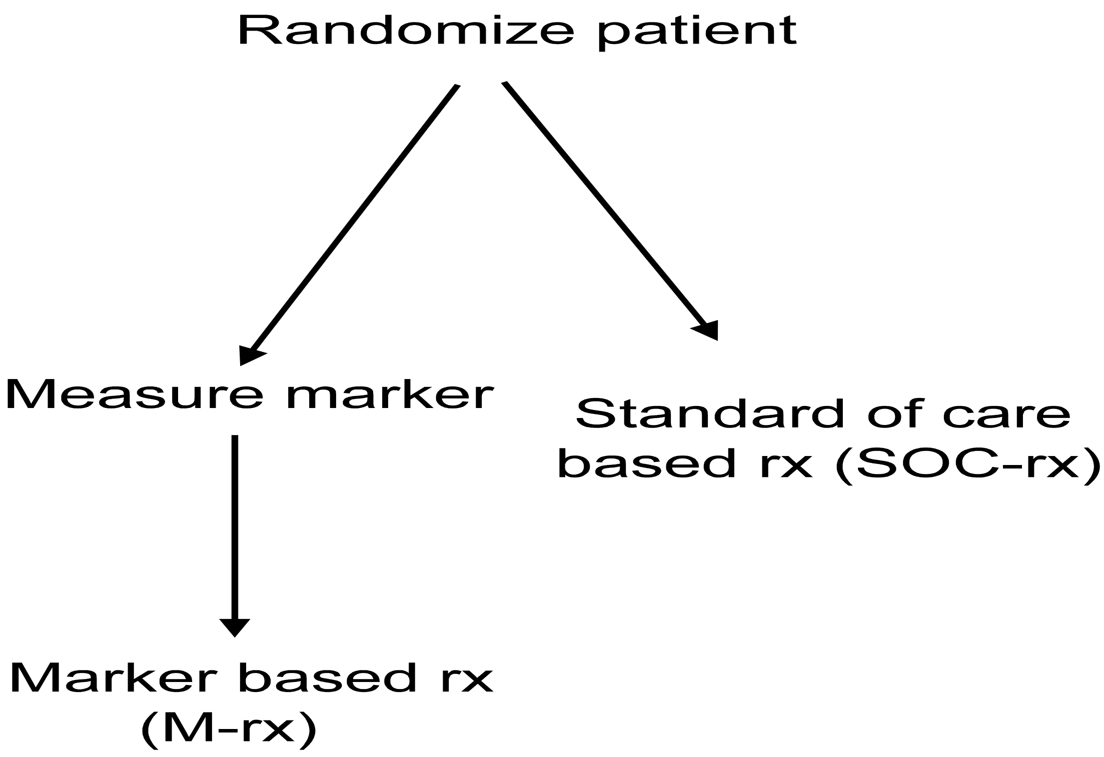
- Developmental studies should develop a completely specified classifier
- Developmental studies are analogous to phase 2 therapeutic trials
- Developmental studies should provide an unbiased estimate of predictive accuracy
  - Statistical significance of association between prediction and outcome is not the same as predictive accuracy
- Developmental studies should estimate to what extent predictive accuracy is greater than that achievable with standard prognostic factors

What is the objective of an independent validation clinical trial?

- (a) to see whether the same set of genes are prognostic with independent data?
- (b) To see whether the classification system can predict outcome?
- (c) To see whether use of the classification system for selecting treatment results in clinical benefit compared to not using it?
- (d) To evaluate clinical benefit under conditions that simulate broad application?

Confirmatory Trials for Evaluating a Classifier of Good Risk with Standard Therapy

- Randomize patients to use or non-use of classifier in treatment selection



Randomize patient

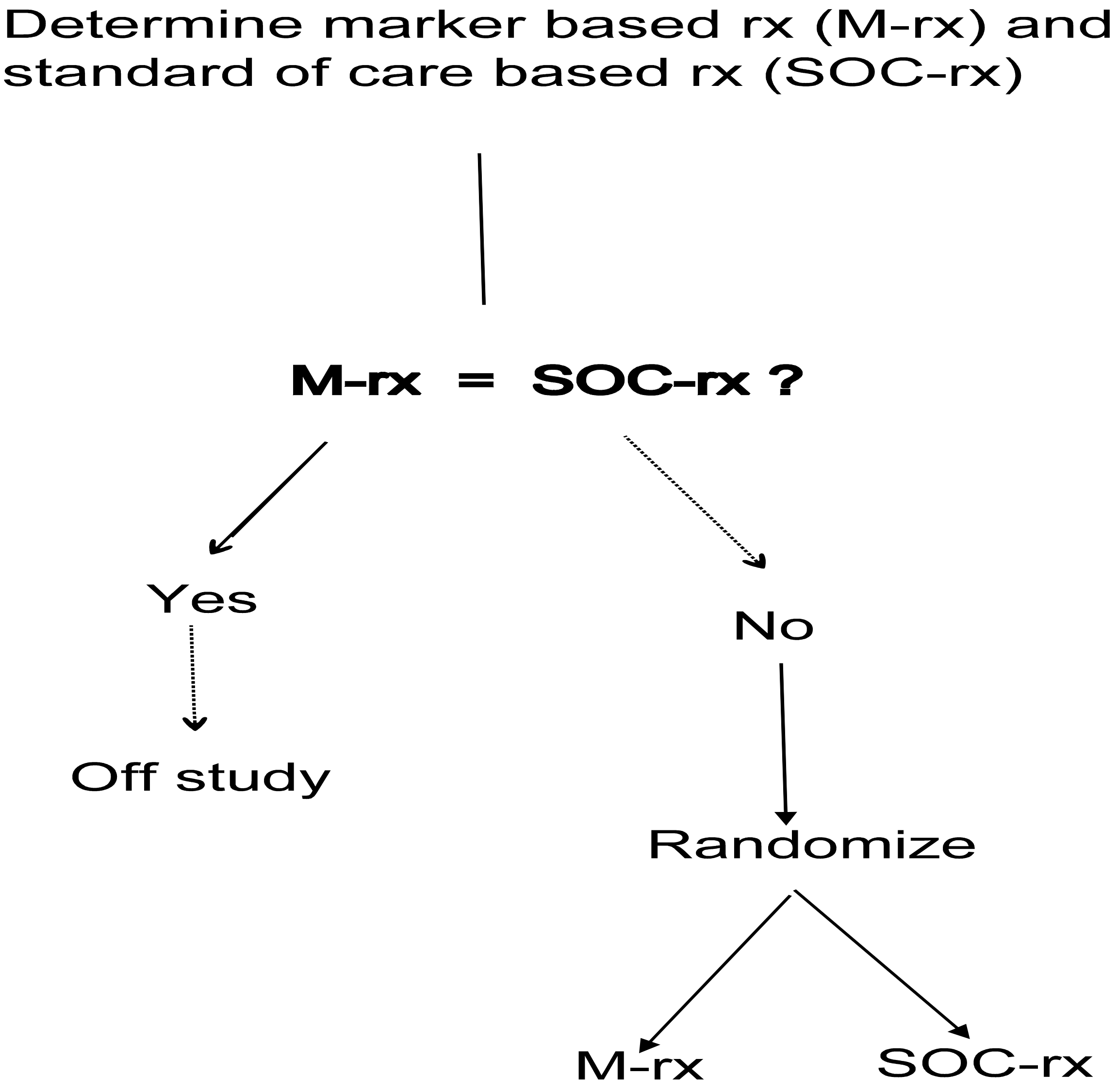
Measure marker

Standard of care based rx (SOC-rx)

Marker based rx (M-rx)

Confirmatory Trials for Evaluating a Classifier of Good Risk with Standard Therapy

- Measure classifier for all patients and randomize only those for whom classifier determined threapy differs form standard therapy



Determine marker based rx (M-rx) and standard of care based rx (SOC-rx)

M-rx = SOC-rx ?

Yes

Off study

No

Randomize

M-rx

SOC-rx

Key Steps in Development and Validation of Therapeutically Relevant Genomic Classifiers

- Develop classifier for addressing a specific important therapeutic decision:
  - Patients sufficiently homogeneous and receiving uniform treatment so that results are therapeutically relevant.
  - Treatment options and costs of mis-classification such that a classifier is likely to be used
- Perform internal validation of classifier to assess whether it appears sufficiently accurate relative to standard prognostic factors that it is worth further development
- Translate classifier to platform that would be used for broad clinical application
- Demonstrate that the classifier is reproducible
- Independent validation of the completely specified classifier on a prospectively planned study