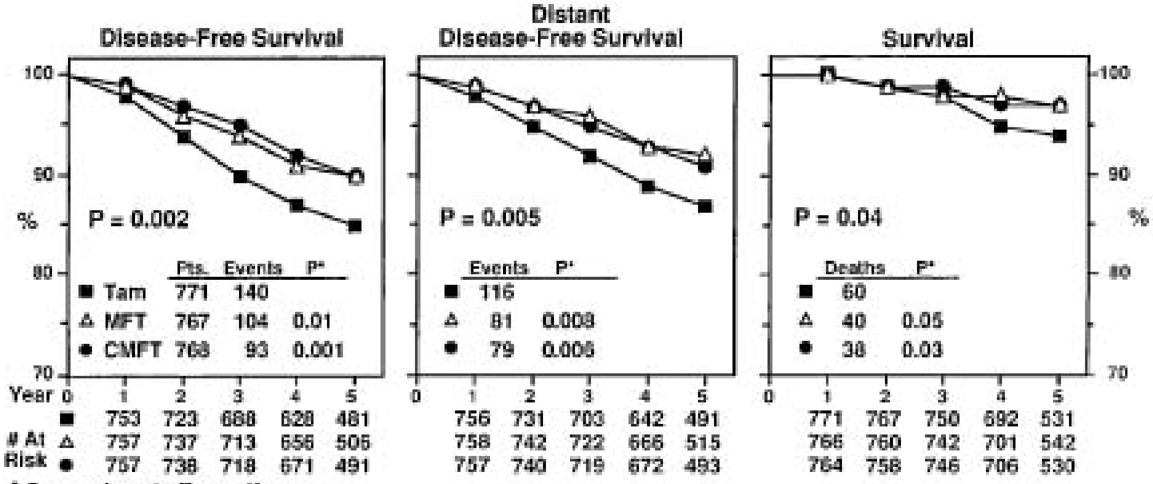
A Roadmap for Developing & Validating Therapeutically Relevant Genomic Classifiers

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

Oncology Needs

- Better tools for utilization of existing treatments

- Better methods for developing new treatments and diagnostics for targeting them to the right patients



# At Risk	Year 1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
■	753	723	688	628	481	756	731	703	642	491	771	767	750	692	531
△	757	737	713	656	506	758	742	722	666	515	766	760	742	701	542
●	757	738	718	671	491	757	740	719	672	493	764	758	746	706	530

* Comparison to Tamoxifen

RELATIVE RISK (95% CONFIDENCE INTERVAL)

MFT/Tam	0.72 (0.56-0.93)	0.68 (0.51-0.90)	0.67 (0.45-0.99)
CMFT/Tam	0.65 (0.50-0.84)	0.67 (0.50-0.89)	0.64 (0.42-0.95)

- Molecularly targeted drugs may benefit a relatively small population of patients with a given primary site/stage of disease
  - Iressa
  - Herceptin

Uses of “Biomarkers”

- Surrogate endpoints
  - A measurement made on a patient before, during and after treatment to determine whether the treatment is working
- Prognostic factor
  - A measurement made before treatment that correlates with outcome, usually for a heterogeneous set of patients
- Predictive factors
  - A measurement made before treatment to predict whether a particular treatment is likely to be beneficial

Surrogate Endpoints

- It is difficult to properly validate a biomarker as a surrogate for clinical outcome.
    - It requires a series of randomized trials with both the candidate biomarker and clinical outcome measured and then demonstration that treatment vs control conclusions for the surrogate are consistent with treatment vs control conclusions for clinical outcome

Surrogate Endpoints

- It is often more difficult to properly “validate” a surrogate than to use the clinical endpoint in phase III trials

Prognostic Factors

- Many prognostic factor studies utilize “convenience samples” of patients that are heterogeneous with regard to disease extent and treatment. The results are often of uncertain relevance for therapeutic decision making.

Pusztai et al. The Oncologist 8:252-8, 2003

- 939 articles on “prognostic markers” or “prognostic factors” in breast cancer in past 20 years
- ASCO guidelines only recommend routine testing for ER, PR and HER-2 in breast cancer
- “With the exception of ER or progesterone receptor expression and HER-2 gene amplification, there are no clinically useful molecular predictors of response to any form of anticancer therapy.”

Predictive Diagnostic Classifiers

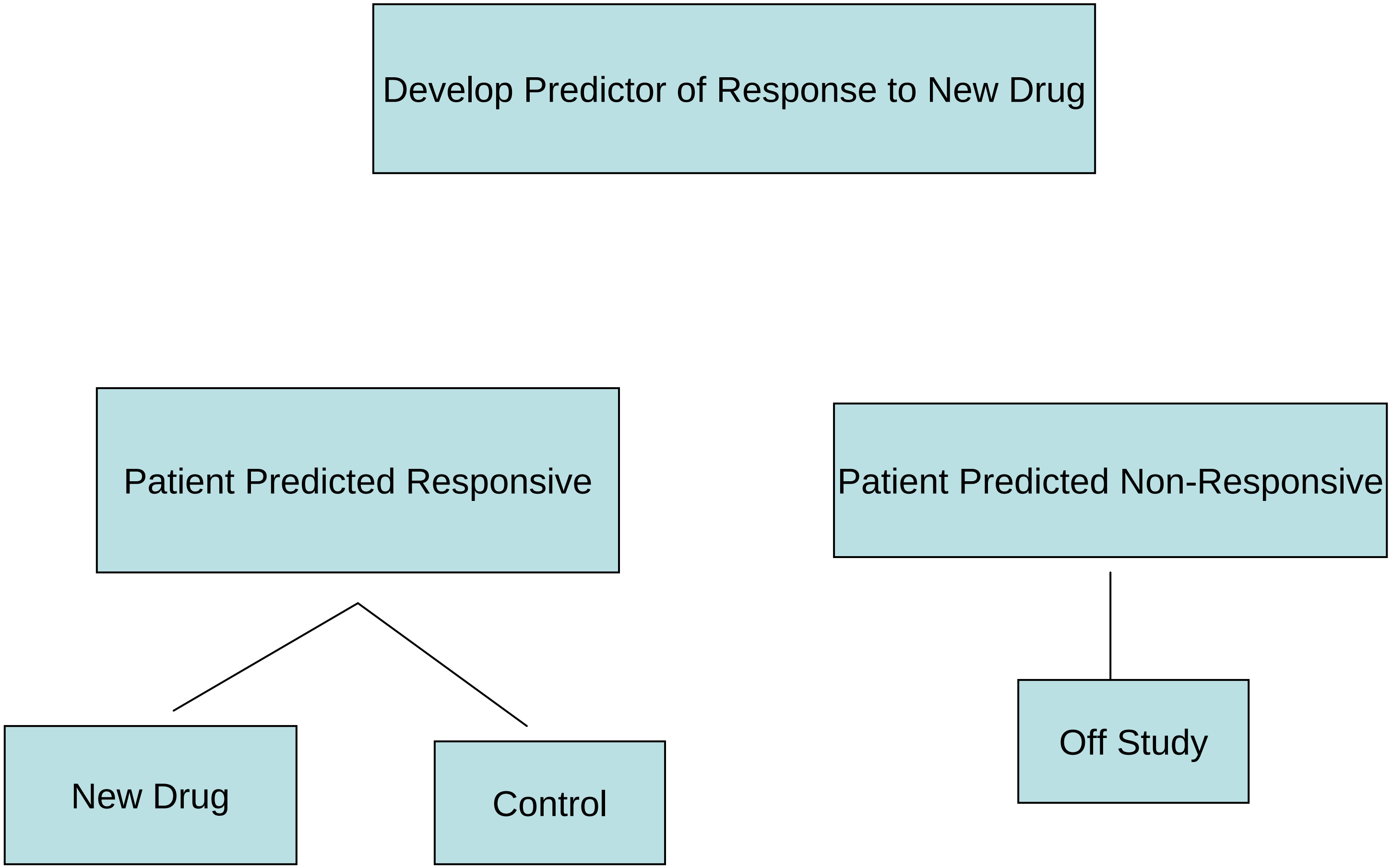
- In new drug development
- For effective utilization of widely available therapies

- In new drug development, the role of a classifier is to select a target population for treatment
  - The focus should be on evaluating the new drug, not on validating the classifier

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

Developmental Strategy (I)

- **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

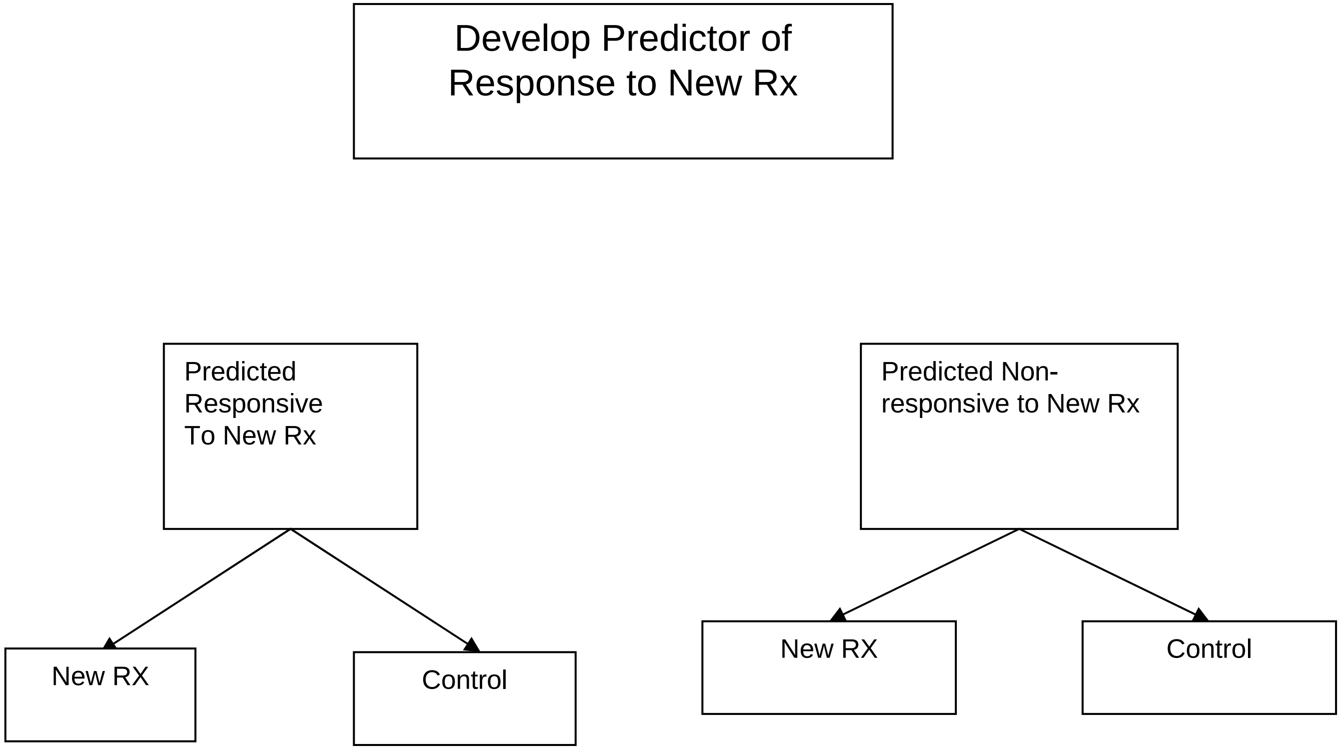
Randomized Ratio

$n_{untargeted}/n_{targeted}$

Proportion Assay Positive	No Treatment Benefit for Assay Negative Patients	Treatment Benefit for Assay Negative Patients is Half That for Assay Positive Patients
0.75	1.78	1.31
0.5	4	1.78
0.25	16	2.56

- For Herceptin, even a relatively poor assay enabled conduct of a targeted phase III trial which was crucial for establishing effectiveness
- Recent results with Herceptin in early stage breast cancer show dramatic benefits for patients selected to express Her-2

Developmental Strategy (II)



Develop Predictor of Response to New Rx

Predicted Responsive To New Rx

New RX

Control

Predicted Non-responsive to New Rx

New RX

Control

Developmental Strategy (II)

- Do not use the diagnostic to restrict eligibility, but to structure a prospective analysis plan.
- Compare the new drug to the control overall for all patients ignoring the classifier.
  - If $p_{overall} \leq 0.04$ claim effectiveness for the eligible population as a whole
- Otherwise perform a single subset analysis evaluating the new drug in the classifier + patients
  - If $p_{subset} \leq 0.01$ claim effectiveness for the classifier + patients.

Key Features of Design (II)

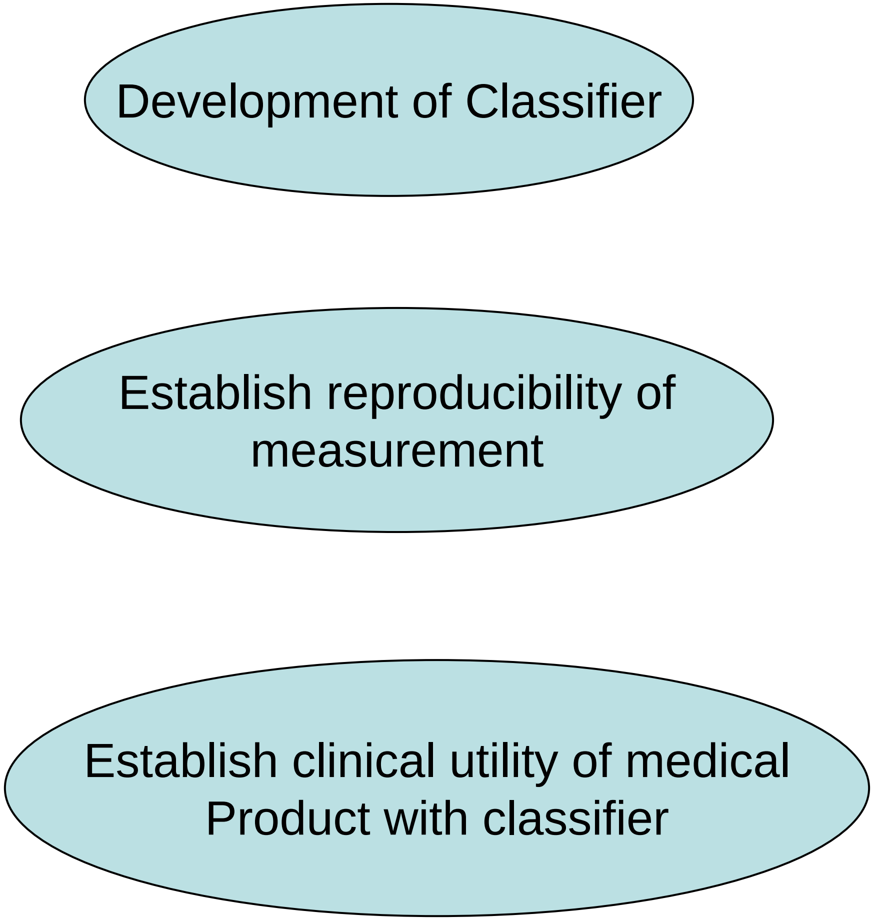
- The purpose of the RCT is to evaluate treatment T vs C overall and for the pre-defined subset; not to modify or refine the classifier

Key Features of Design (II)

- Pre-specified analysis plan
- Single pre-defined subset
- Overall study type I error of 0.05 is split between overall test and subset test
- Saying that the study should be “stratified” is not sufficient
  - It doesn’t matter whether randomization is stratified except that it helps ensure that all patients have specimens available to assay for classification

The Roadmap

1. Develop a completely specified genomic classifier of the patients likely to benefit from a new medical product
2. Establish reproducibility of measurement of the classifier
3. Use the completely specified classifier to design and analyze a new clinical trial to evaluate effectiveness of the new treatment with a pre-defined analysis plan.



Development of Classifier

Establish reproducibility of measurement

Establish clinical utility of medical Product with classifier

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

Use of Archived Samples

- Archived samples from a conventional non-targeted “negative” clinical trial can be used to develop a binary classifier of a subset thought to benefit from treatment.
- That subset hypothesis should be tested in a separate clinical trial
  - Prospective targeted type (I) trial
  - Prospective type (II) trial
  - Analysis of archived specimens from a second previously conducted clinical trial to identify classifier positive patients

Development of Genomic Classifiers

- Single gene or protein based on knowledge of therapeutic target or

- Empirically determined based on correlating gene expression to patient outcome after treatment or

- Single gene or protein culled from set of candidate genes identified based on imperfect knowledge of therapeutic target

Development of Genomic Classifiers

- During phase I/II development or

- After failed phase III trial using archived specimens.

- Adaptively during early portion of phase III trial.

Adaptive Signature Design

An adaptive design for generating and prospectively testing a gene expression signature for sensitive patients

Boris Freidlin and Richard Simon

Clinical Cancer Research 11:7872-8, 2005

Adaptive Signature Design
End of Trial Analysis

- Compare E to C for **all patients** at significance level 0.04
  - If overall H_0 is rejected, then claim effectiveness of E for eligible patients
  - Otherwise

- Otherwise:
  - Using only the first half of patients accrued during the trial, develop a binary classifier that predicts the subset of patients most likely to benefit from the new treatment E compared to control C
  - Compare E to C for patients accrued in second stage who are predicted responsive to E based on classifier
    - Perform test at significance level 0.01
    - If H_0 is rejected, claim effectiveness of E for subset defined by classifier

Use of DNA Microarray Expression Profiling

- For settings where you don’t know how to identify the patients likely to be responsive to the new treatment based on its mechanism of action
- Only pre-treatment specimens are needed
- Expression profiling should be used to identify informative genes and form a binary classifier

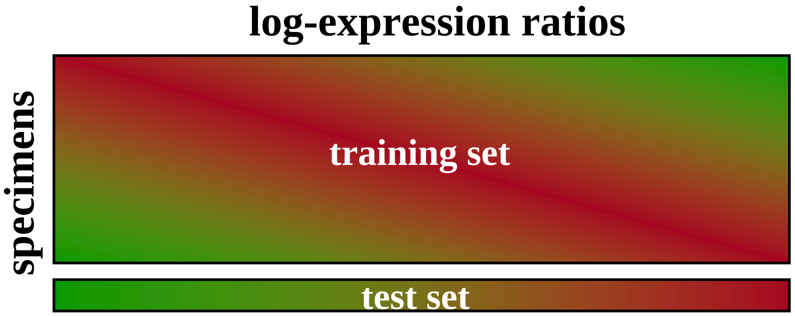
Internal Validation

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy

Split-Sample Evaluation

- Divide cases into a training set and a testing set
- Do not permit any access to the testing set until a single fully specified classification model is developed on the training set

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.

Limitations of Expression Profiling Studies for Prediction Problems

- Over-use and misleading use of cluster analysis
- Test sets too small
- Improperly applied cross-validation
- Failure to demonstrate that prediction accuracy is better than that achievable with standard clinical and histologic variables

Limitations to Internal Validation

- Sample handling and assay conduct are performed under controlled conditions that do not incorporate real world sources of variability
- Developmental studies are generally small
- Predictive accuracy is generally not clinical utility

Studies Developing Gene Expression Profile Classifiers Should be Viewed as Analogous to Phase II Trials Requiring Phase III Validation

External Validation

- From different clinical centers
- Specimens assayed at different time from training data
- Reproducibility of assay for individual tumors demonstrated to clinical reference laboratory standards
- Positive and negative samples collected in the same way
- Study sufficiently large to give precise estimates of sensitivity and specificity of the classifier
- Study addresses clinical utility of using the genomic classifier compared to using standard practice guidelines

Adequate External Validation Studies are Rarely Performed

- They are expensive, require multi-center cooperation and large sample sizes
- The financial incentives for developing and validating diagnostics for optimizing use of available treatments are limited

Conclusions

- New technology and biological knowledge are sufficiently mature to support the prediction of which patients will benefit from new treatments and to personalize the use of existing treatments
- Targeting treatment can dramatically improve the efficiency of clinical trials, benefit patients and help control costs of medical care

Conclusions

- Much of the conventional wisdom about how to develop and utilize predictive classifiers is flawed
- Leadership is needed to harness currently available technology to support the development and validation of diagnostics to transit from 20th century correlative science to 21st century predictive medicine

Conclusions

- Prospectively specified analysis plans for phase III data are essential to achieve reliable results
  - Biomarker analysis does not mean exploratory analysis except in developmental studies
- In some cases, definitive evidence can be achieved from prospective analysis of patients in previously conducted clinical trials with extensive archival of pre-treatment specimens

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), 2005.

Freidlin B and Simon R. Adaptive signature design. Clinical Cancer Research 11:7872-8, 2005.

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, pp 3-15, 2005.

Simon R and Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. The Pharmacogenomics Journal, 2006.