Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. Richard Simon, J Clin Oncol 23:7332-7341

Presented by Deming Mi

7/25/2006

Major reasons for few prognostic factors to be used in clinical practice

- lack of impact or guideline on clinical practice (prognostic factors should be therapeutically relevant, e.g. a classifier for predicting the patients unlikely to respond to a certain therapy may not be widely used if there is no good alternative treatment)
- lack of well defined patients with respect to disease status, therapy etc.
- lack of independent validation of prognostic markers
- lack of detailed analysis planning and clear define of participants and end points
- lack of reproducibility study

Key Steps in Development and Validation of Therapeutically Relevant Genomic Classifiers

- **Step 1**: Develop classifier for addressing a specific important therapeutic decision
  - Patients are sufficiently homogeneous and receiving uniform treatment so that results are therapeutically relevant
  - Treatment options and costs of mis-classification are such that a classifier is likely to be used
- **Step 2**: Perform internal validation of classifier to assess whether it appears sufficiently accurate relative to standard prognostic factors that it is worth further development
- **Step 3**: Translate classifier to platform that would be used for broad clinical application
- **Step 4**: Demonstrate that the classifier is reproducible
- **Step 5**: Independent validation of the completely specified classifier on a prospectively planned study

Some issues about multigene classifier

- A multigene expression signature classifier is a function that provides a classification of a tumor based on the expression levels of the component genes.
- genes regulating the same pathway are correlated
- simultaneous change of genes involved in regulating the same signal pathway may be significant biologically, but not necessarily significant statistically
- stringent statistical criteria used for identifying differential genes results in reduced power and may miss those co-regulated genes
- classifier external validation validates the usefulness of classifier in clinical practice, not the development of classifier

Step 1: Develop classifier for addressing a specific important therapeutic decision

Steps in classifier development

- determine which features should be included as components, e.g. select the top 10 most differential genes that best discriminate between two class of subjects
- what mathematical form to use for combining the values of the different features, e.g. $l(\underline{x}) = \sum_{i \in F} w_i x_i$
- what cutoff values to use for converting a continuous function into a discrete classification

What kinds of classifiers are useful

- Fisher linear discriminant analysis (FLDA)
- Maximum likelihood discriminant rules (DQDA, DLDA)
- Weighted voting method of Golub (Golub)
- Nearest neighbor classifiers (NN)
- Classification and regression trees (CART)
- CART with perturbed learning sets generated by bagging (Bag CART)
- CART with perturbed learning sets generated by sampling from MVN (MVN CART)
- CART with perturbed learning sets generated by convex pseudo-data (CPD CART)
- CART with perturbed learning sets generated by boosting (Boost CART)

What did Dudoit et al do?

- Leukemia dataset described by Golub 1999
- 47 cases of ALL and 25 cases of AML
- p = 40 selected genes (genes with largest BSS/WSS)
- 2:1 of training to test set ratio
- repeat entire procedure N = 150 time
- compute test set error rate for each classifier
- compute observation-wise error rate

Dudoit S. J Am Stat Assoc 97:77-87, 2002

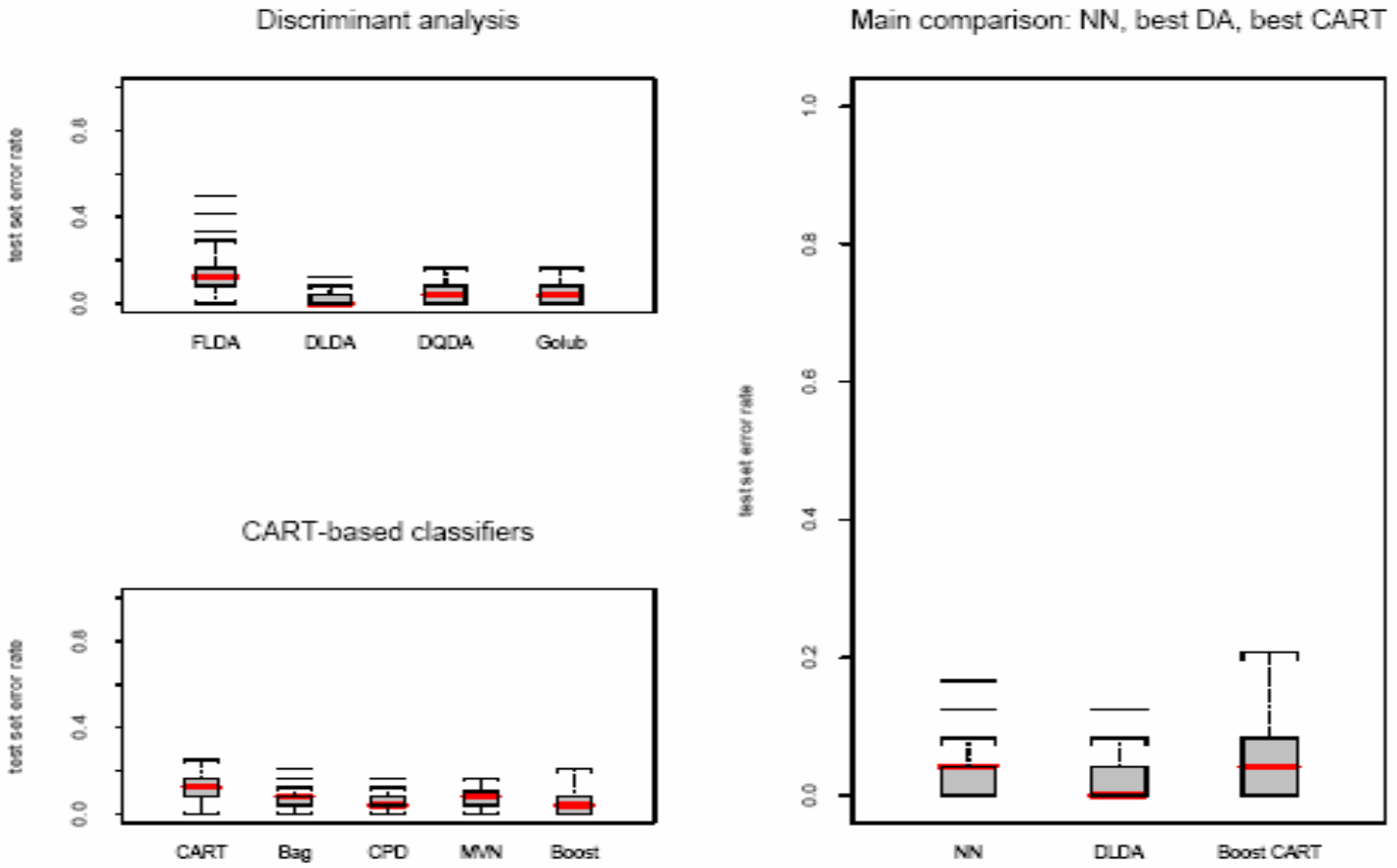


Figure 5: *Leukemia data, two classes - Test set error rates.* Boxplots of test set error rates for classifiers built using the $p = 40$ genes with the largest BSS/WSS; $N = 150$ LS/TS runs for $2:1$ sampling scheme.

Dudoit S. J Am Stat Assoc 97:77-87, 2002

•Nearest Neighbor (NN) and Diagonal Linear Discriminant Analysis (DLDA) classifiers have the smallest mean error rate
•Fisher's Linear Discriminant Analysis (FLDA) has the highest error rate

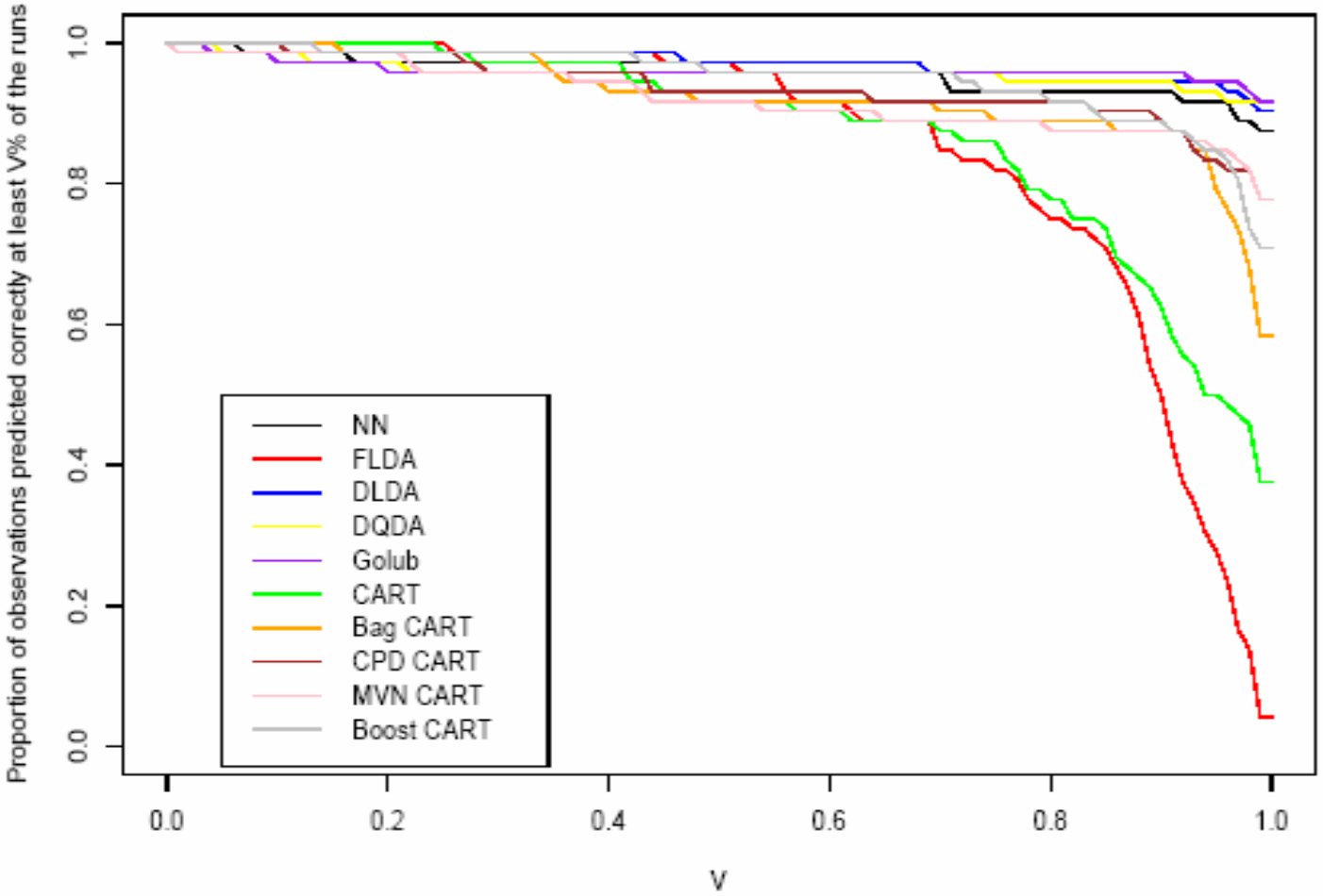


Figure 11: *Leukemia data, two classes - observation-wise error rates.* Fraction of observations predicted correctly at least $V\%$ of the time (out of the runs for which a given observation belonged to the test set) for $p = 40$ genes; $N = 150$ LS/TS runs for $2:1$ scheme.

Dudoit S. J Am Stat Assoc 97:77-87, 2002

•NN and DLDA perform as well or better than more complex classifiers

•FLDA perform poorly

- The fact that number of cases is relatively smaller compared to the number of genes prevents the more complex nonlinear classifier to perform superiorly than the simplest linear classifier
- In a small data set (this is probably true for most cases), there is not sufficient information to effectively utilize complex classifiers

Step 2:Perform internal validation of classifier to assess whether it appears sufficiently accurate relative to standard prognostic factors that it is worth further development

- Purpose of internal validation: provide a preliminary estimate of the predictive power of the classifier
- Danger of using error rate or predictive accuracy from development study based on training set:
  - overfitting almost always occur, always possible to find classifiers with high predictive accuracy even there is no relationship between expression of any of the genes and outcome;
  - fit random variation or noise in original data that do no represent true relationships that hold for independent data

Cross Validation

- Holdout method or split-sample validation
  - withholding a substantial proportion of the sample from training set may considerably reduce the performance of the prediction rule
  - not an efficient way to use the available data
- K-fold cross validation
- Leave-one-out cross validation (K-fold cross validation with K = N)

What does cross validation validate

- Three steps in class prediction
  - selection of informative genes
  - computation of weights for selected informative genes
  - creation of a prediction rule
- It is important that all three steps undergo the cross-validation procedure
- Cross-validated prediction error is an estimate of the prediction error associated with application of the algorithm for model building
- Failure to cross-validate all steps may lead to substantial bias in the estimated error rate

What did Simon et al do?

- Each simulated data set is composed of 20 gene expression profiles, each profile consisting of 6000 gene expression measurements
- For each of the 20 profiles, the 6000 genes expression measurements are independent and identically distributed from the standard normal distribution (mu=0, var=1)
- Randomly assign 10 profiles to class 1 and the other 10 to class 2
- Generate 2000 such simulated data sets

Simon R, J Natl Cancer Inst 95:14-18, 2003

Steps in class prediction study

- Gene selection (10 most differentially expressed genes based on two-sample t-test)
- Computation of a weight for each gene (univariate two-sample t-statistic) $l(\underline{x}) = \sum_{i \in F} w_i x_i$
- Computation of prediction rule (classification threshold)

Resubstitution (no cross-validation)

- Gene selection (10 most differentially expressed genes based on two-sample t-test)
- Computation of a weight for each gene (univariate two-sample t-statistic)
- Computation of prediction rule (classification threshold)

Cross-validation after gene selection

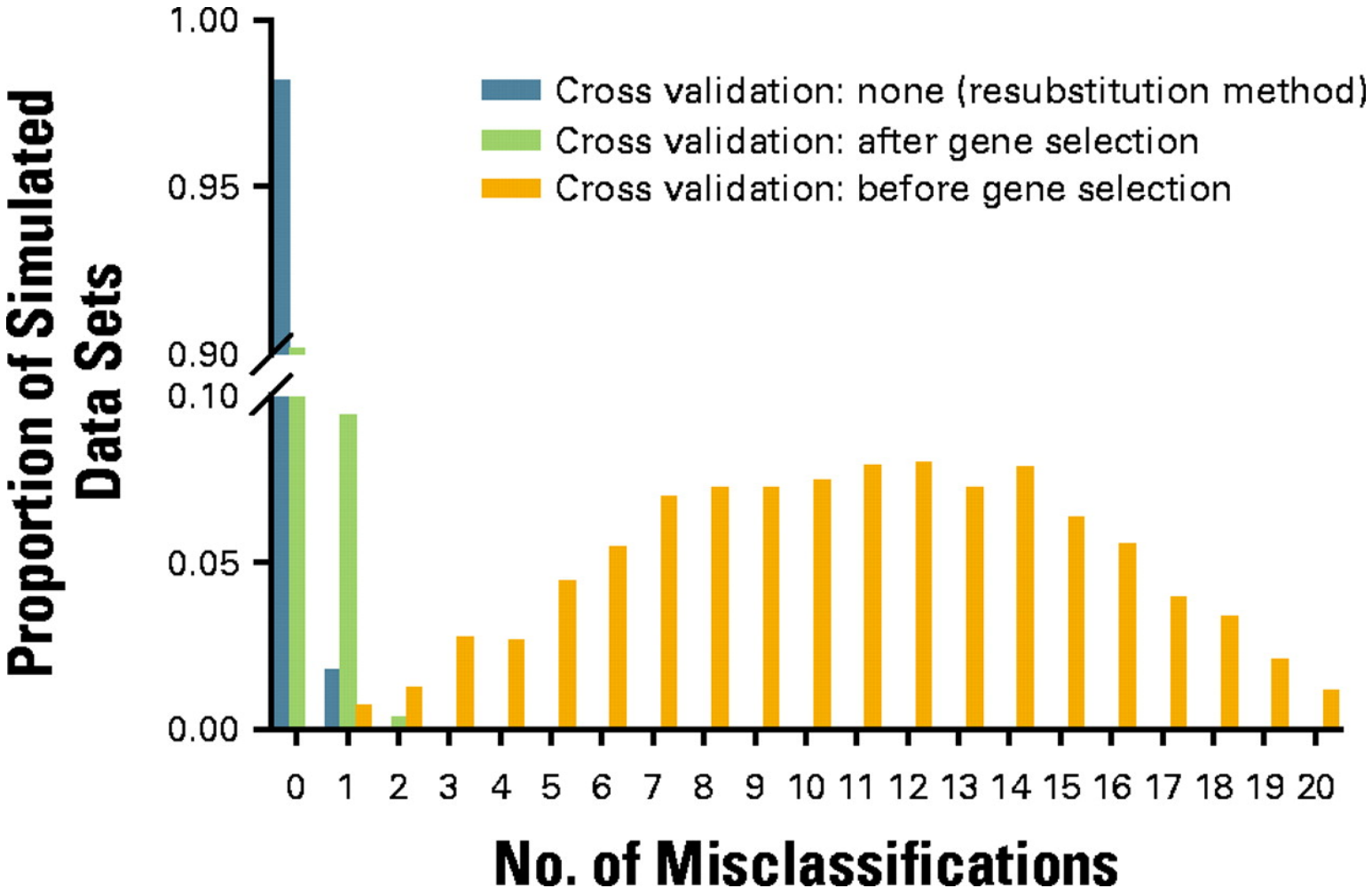
- Gene selection (10 most differentially expressed genes based on two-sample t-test)

- Computation of a weight for each gene (univariate two-sample t-statistic)
- Computation of prediction rule (classification threshold)

Cross-validation prior to gene selection

- Gene selection (10 most differentially expressed genes based on two-sample t-test)
- Computation of a weight for each gene (univariate two-sample t-statistic)
- Computation of prediction rule (classification threshold)

Fig 1. The effect of various levels of cross validation on the estimated error rate of a predictor



Simon, R. J Clin Oncol; 23:7332-7341 2005

- The improperly cross-validated method results in seriously biased underestimate of the error rate, probably largely due to overfitting the predictor to the specific dataset

How to report the error rate

- Report properly cross-validated error rate
- Report error rate derived from sufficiently large independent validation set
- Report error rate from validation on all the classes for which the classifier was developed
- Report confidence interval or statistical significance of the error rate
  - especially when the test set is small
  - the point estimate of error rate may be small while the CI is too wide and cover 0.5
- Consider unclassified specimen
  - Unclassified specimens are those that could not confidently be assigned to any of the examined classes based on the classifier
  - Simply ignoring the unclassified specimens gives an overstated performance of classifier

Does your classifier perform better than standard prognostic factors?

- the new classifier needs to outperform or add predictive power to the existing prognostic methods in order to justify the money and time invested on the external validation in a trial of much larger scale

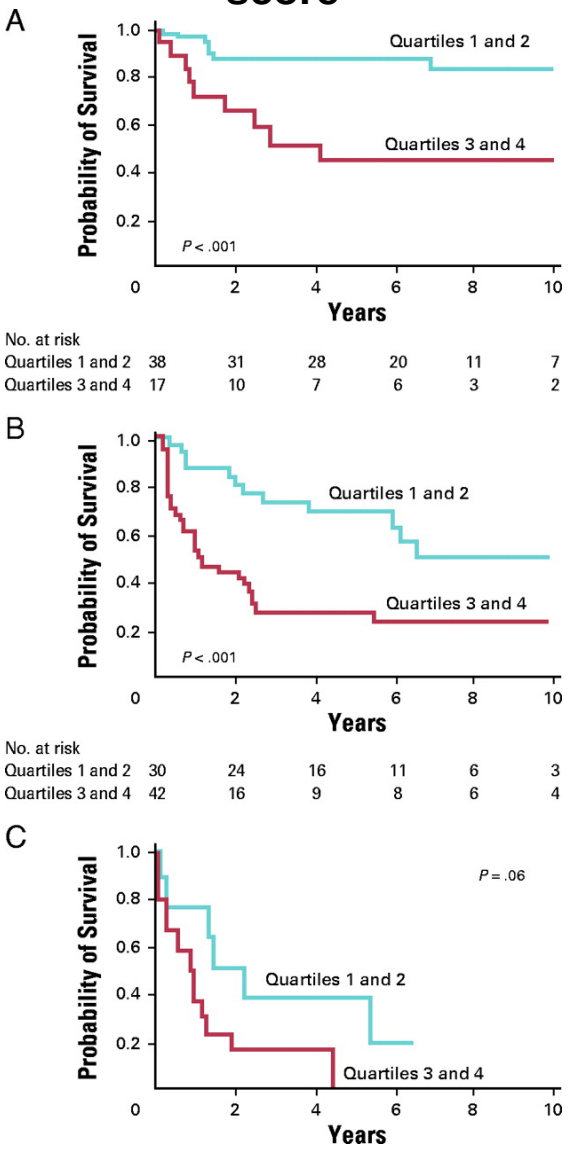
Rosenwald A et al, The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346:1937-1947, 2002

- 670 of 7399 microarray features were significantly associated with a good or a bad survival outcome in the training set (N=160)
- Selected 16 genes that were highly variable in expression out of the 670 differential genes in order to minimize the number of genes in the outcome predictor
- Categorized the 16 genes based on their physiological function.
- outcome-predictor score = (0.241 x the average value of the proliferation signature) + (0.310 x value for *BMP6*) – (0.290 x the average value of the germinal-center B-cell signature) – (0.311 x the average value of the major-histocompatibility-complex [MHC] class II signature) – (0.249 x the average value of the lymph-node signature)
- High score indicates a poor outcome.

Will the gene-expression–based outcome-predictor score add more predictive accuracy to the standard International Prognostic Index (IPI)?

Rosenwald used a validation set of N=80 to see if the predictor score can add predictive power to IPI.

Fig 2. Survival curves for diffuse large-B-cell lymphoma patients by gene expression classifier stratified by three levels of International Prognostic Index (IPI) score: (A) IPI scores 0-1; (B) IPI scores 2-3; (C) IPI scores 4-5.Four prognostic classes were defined based on gene expression risk score



Simon, R. J Clin Oncol; 23:7332-7341 2005

Step 3 & 4: Translate classifier to platform that would be used for broad clinical application and demonstrate that the classifier is reproducible

- RT-PCR can be performed on formalin-fixed paraffin-embedded (FFPE) tissue
- Standardization of assay protocols is crucial to achieve satisfactory inter- and intra-laboratory reproducibility

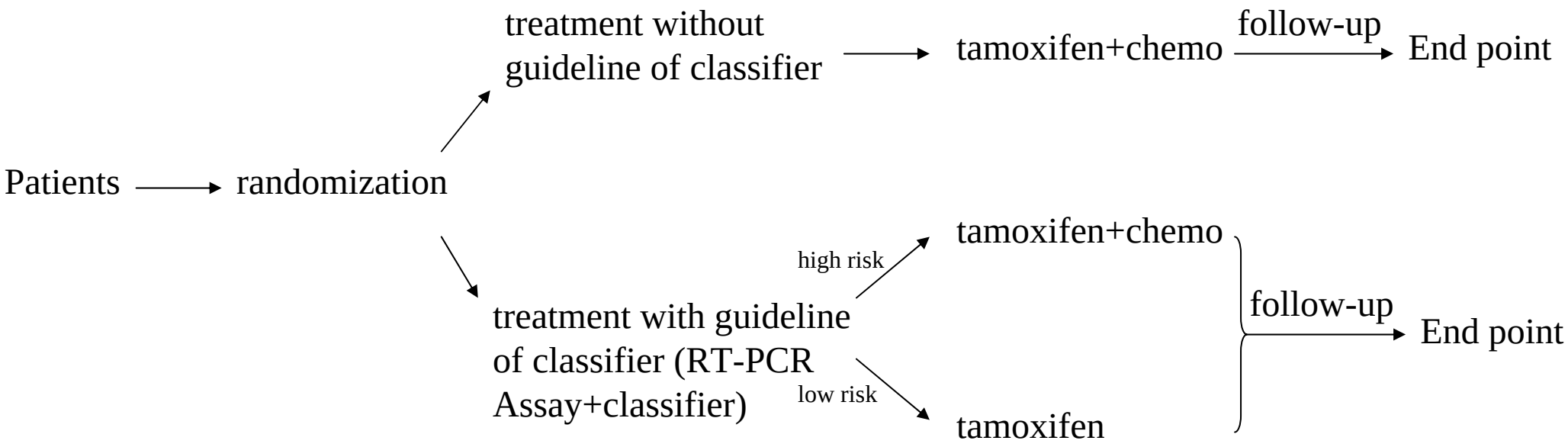
Step 5: Independent external validation of the completely specified classifier on a prospectively planned study

- The objective is to determine whether use of a completely specified diagnostic classifier for therapeutic decision making in a defined clinical context results in patient benefit
- The objective is not to see if the same genes are prognostic or if the same classifier is obtained

Example

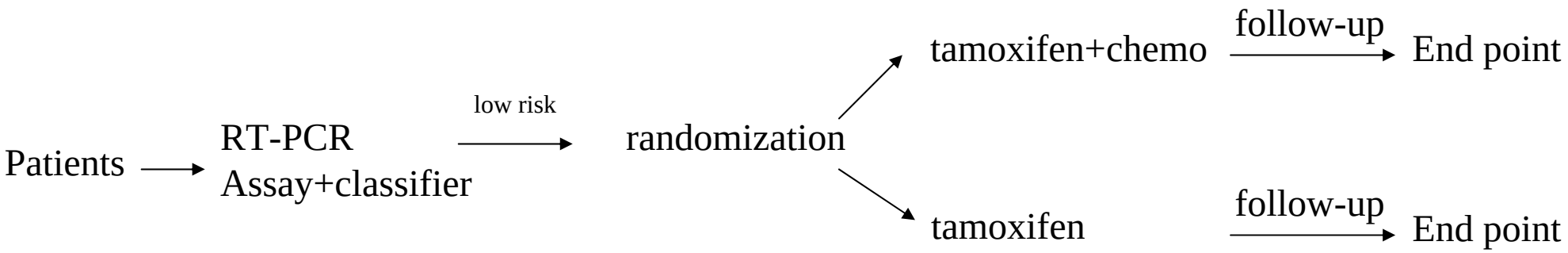
- Paik S, Shak S, Tang G, et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817-2826, 2004
- Constructed a recurrence-score algorithm based on training set of N=447 patients
- Classify patients into low risk (recurrence-score<18), intermediate risk (18≤recurrence-score<31), and high risk (31≤recurrence-score)
- Want to test the strategy of withholding cytotoxic chemotherapy form the subset of patients classified as low risk

Design 1



- Patients which are categorized as high risk will receive the same treatment (tamoxifen+chemo), either way they are assigned
- Require huge sample size, especially when high-risk patients account for the majority of the cohort

Design 2



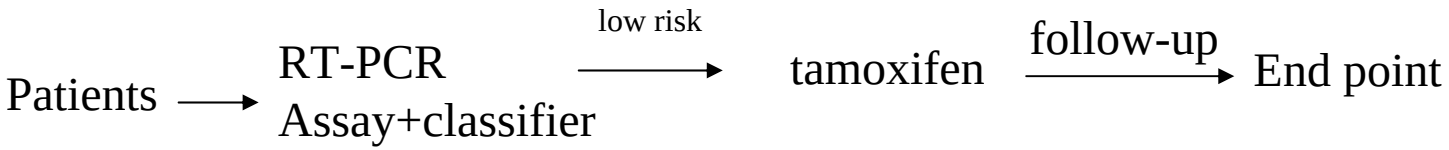
Pros: Require less patients and more efficient than the previous design assigning all patients

Cons:

•Recurrence rate is low for low-risk patients

•With the fact that the recurrence rate is low, it still requires many patients to be able to detect small difference in order to reject null hypothesis and claim bioequivalence.

Design 3



- Single-arm
- Require the smallest number of patients
- If after long follow-up, the recurrence rate is very low, then the classifier is considered validated for providing clinical benefit because it enable the identification of patients whose prognosis was so good with tamoxifen monotherapy that they could be spared the cytotoxic chemotherapy.

Table 1. Kaplan–Meier Estimates of the Rate of Distant Recurrence at 10 Years, According to Recurrence-Score Risk Categories.*

Risk Category	Percentage of Patients	Rate of Distant Recurrence at 10 Yr (95% CI)†
		percent
Low	51	6.8 (4.0–9.6)
Intermediate	22	14.3 (8.3–20.3)
High	27	30.5 (23.6–37.4)‡

* A low risk was defined as a recurrence score of less than 18, an intermediate risk as a score of 18 or higher but less than 31, and a high risk as a score of 31 or higher.
† CI denotes confidence interval.
‡ $P<0.001$ for the comparison with the low-risk category.

Paik S, et al N Engl J Med 351:2817-2826, 2004

- Use of a genomic classifier for focusing a clinical trial on a population which is more likely to benefit can reduce the required sample size dramatically
- When long term follow-up is difficult, can conduct a prospectively planned validation using archived specimens from patient treated in a previously conducted prospective multicenter clinical trial
- Study protocol should be developed prospectively

- In Paik S, et al 2004, they stated: “The prospectively defined assay methods and end points were finalized in a protocol signed on August 27, 2003. RT-PCR analysis was initiated on September 5, 2003, and RT-PCR data were transferred to the NSABP for analysis on September 29, 2003. “