Development of Biomarker Signatures from High-Dimensional Data

V. Devanarayan, Ph.D.
Abbott Laboratories, USA
viswanath.devanarayan@abbott.com

Non-Clinical Statistics Conference
Lyon, France, September 27-29, 2010

Outline

Overview about biomarkers, signatures, etc.

Biomarker Signature Development

- Process
- Filtering/Selection
- Subset derivation
- Cross-Validation
- Batch-Effect normalization
- Targeted focus on biological pathways

Summary

Typical uses of biomarkers in pharmaceutical drug development

During early phase studies, need a quicker read on the efficacy/safety;

- E.g., 6-month change in gene/protein expression that predicts 2-year cognitive decline.

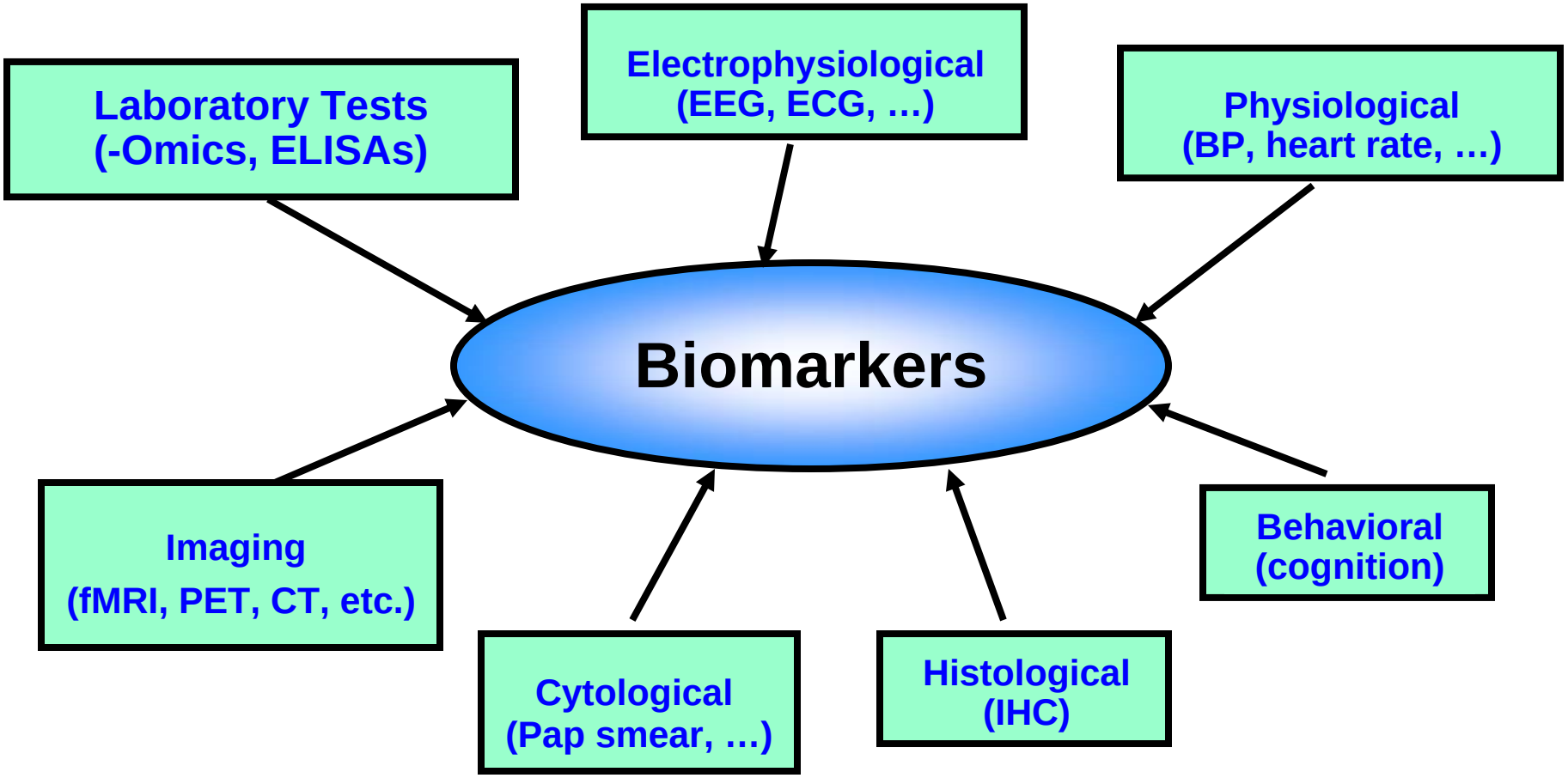
For patient-selection in the clinical trial.

- Biomarkers that accurately diagnose severity of AD, Cancer, etc.
- Biomarkers that provide early read on the disease progression.

For predicting responders & non-responders to a drug, which patients should get higher dose, etc.

For early read on safety/tox such as liver and kidney injury.

Where do these Biomarkers come from?



Biomarker may be a Gene, Protein, Imaging, Clinical measure, etc.

Biomarker Signatures

Biomarker Signature is a combination of one or more markers, when applied to an empirical model or rule, *predicts* an outcome of interest.

- Outcome may be *disease process, drug efficacy, safety*, etc.
- Defined by list of markers & empirical model/rule

May be as complex as

- "25-gene in a tree-based model" for predicting patient response
- "15 proteins in a LDA model" for predicting tumor progression.

or as simple as

- "10 cognitive measures with a decision threshold on their average" for predicting Alzheimer's disease status.
- Decision threshold on a single protein to predict a future endpoint.

Biomarker Signatures (contd.)

Variety of possible data-sets (usually “small n, large p”)

- # of subjects: 10s to 100s (n)
- # of markers: 10s to 100s to 1000s to Millions (p)
  - Genomic arrays
    - CGH Arrays: 150K, 500K, or 2-3 million SNPs
    - mRNA arrays: 20,000 – 50,000 gene probe-sets
    - miRNA arrays: ~500 to 1000 genes
  - Proteomic Arrays: 10-100-1000s of peptides/proteins
  - Imaging: Voxel-level (several 1000s) or “region aggregates (10-100s).
  - Clinical observations: 10s to 100s of clinical measures

Often, multiple sources of biomarker data are available from the same set of subjects (e.g., clinical + genomic).

Biomarker Signatures (contd.)

Typically, signatures include markers from only one source

- gene signature, proteomic signature, imaging signature, etc.

Often appropriate, and possibly more powerful, to include markers from multiple sources.

- *"Extreme" example:* A 12-marker signature for predicting *patient prognosis, disease progression*, etc., may include
  - 3 proteins measured using ELISAs
  - 2 genes; mRNA expression from TaqMan
  - 3 imaging variables (e.g., k-trans for Cancer, vMRI measures for AD)
  - 2 genetic variables (e.g., ApoE for AD)
  - 2 clinical variables (baseline Age, Cognition, etc.)

Usually more practical to expect a combination of 2 to 3 sources (e.g., imaging+proteomic, genetic+clinical, etc.)

Biomarker Signature Development

Predictive inference, Not statistical inference!

Interest is in how well a biomarker signature predicts an outcome of interest, usually at the individual level.

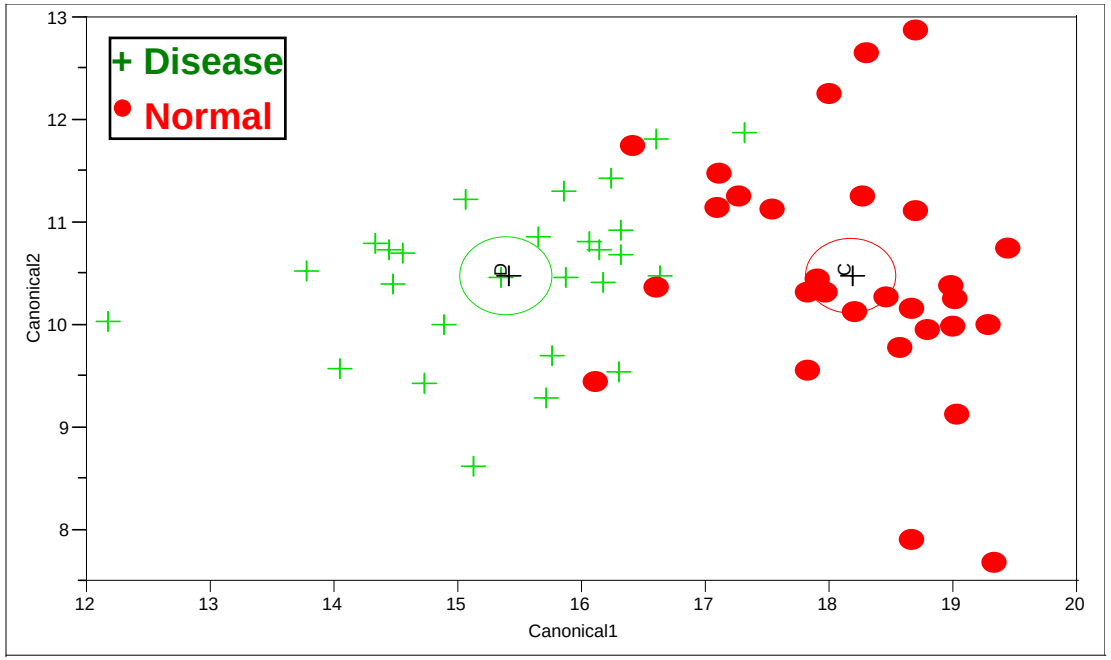
- Typically, "individual" refers to a
  - patient/subject (drug response, disease process, etc.),
  - compound (prediction of compounds likely to toxic), etc.

Need minimal overlap of the individuals between the groups that are compared.

- Not adequate if the groups are different with respect to just their mean biomarker response ($p<0.05$ doesn't mean much!).

Thus the focus is on a *rigorous empirical assessment of the "predictive performance"*, and not just *p* or *q* values.

Composites are usually better than singles



Discriminant Analysis

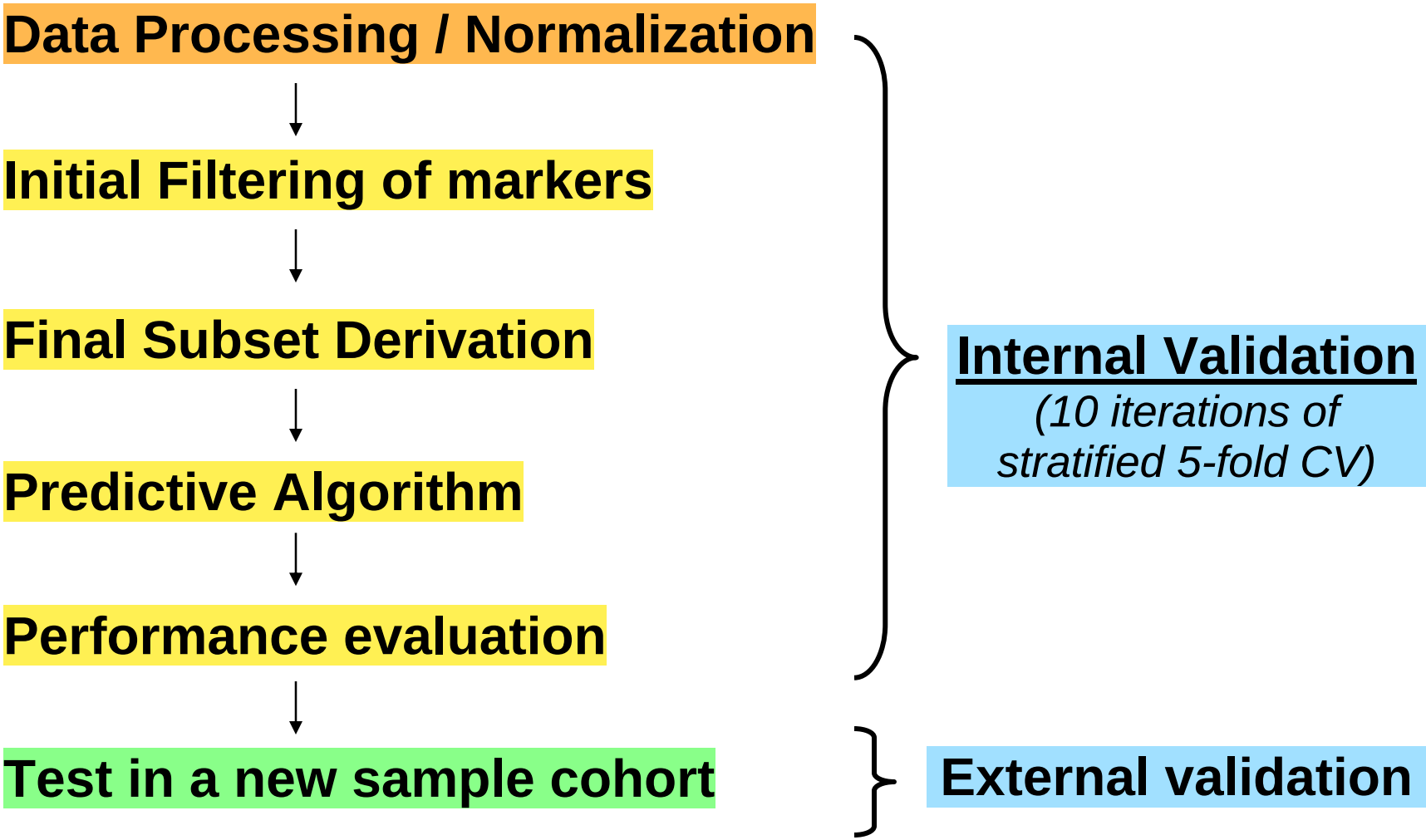
Analysis of a 6-marker panel to discriminate Disease from Normal.

Predictive Accuracy of this composite-marker = 94%

Each marker is statistically significant ($p < 0.05$).

But ***predictive accuracy of each on its own is < 70%.***

Process of biomarker signature development from high-dimensional data



Filtering of markers

Using robust statistical and biological criteria

Statistical criteria

- Univariate/Multivariate methods for selecting Top-X genes.
  - Relative importance scores from a multivariate model on all data
  - SAM *(Significant Analysis of Microarrays; Tusher et al., PNAS, 2001)*
  - Robust, nonparametric or other tests, depending on endpoint/data.
  - In addition, criteria on fold-change & %CV may be used.
- # of markers to filter depends on the context, dataset & model used.
- Small number (25-500) is often adequate, but not always.

Biological criteria

- *Biologically relevant markers from*
  - *disease or target pathway analysis*
  - *Prior internal research, other experience, etc.*
- *Useful to consider initially, even if not statistically significant.*

A disease pathway marker, even if useless on it's own, may be useful in a composite

Case-Study:

A marker, "mk.1", is highly significant on it's own, and provides **75%** predictive accuracy.

Another marker, "mk.X", although **biologically relevant from pathway analysis**, is statistically useless on its own (p>0.05).

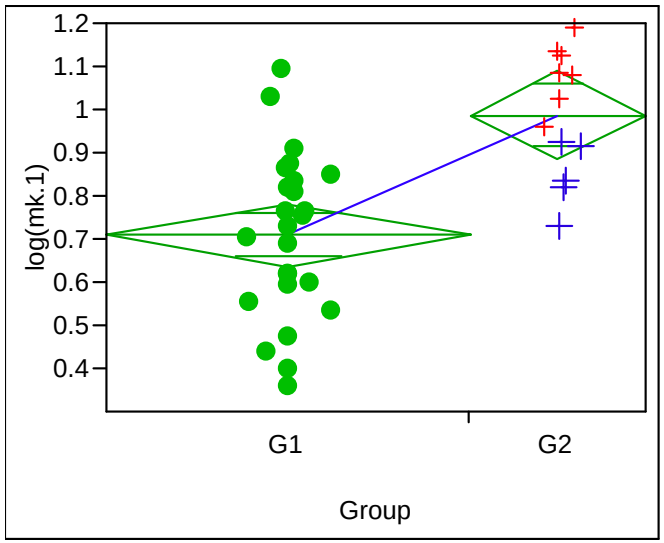
But when mk.X is combined with mk.1, predictive accuracy = **89%**.

Surprised?

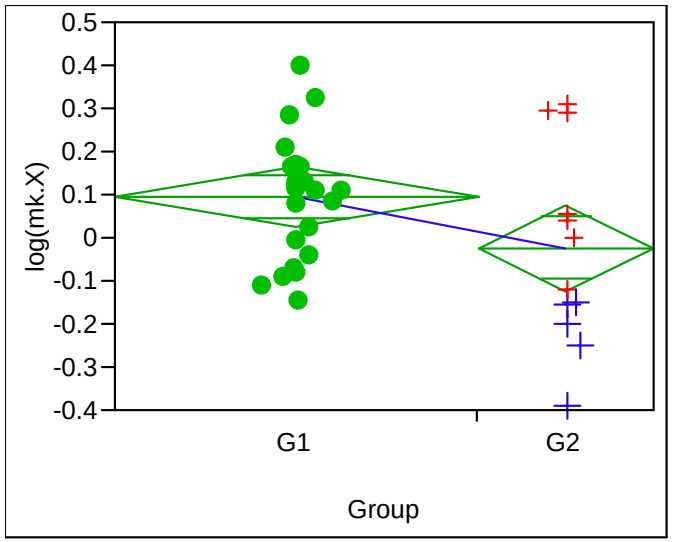
Let's look at a scatter-plot.

A disease pathway marker, useless on it's own, may be useful in a composite: Case Study (contd.)

mk.1



mk.X



Patients that overlap with respect to mk.1 do not overlap in mk.X

Thus when combined into a signature, the overall prediction accuracy increases significantly. This was confirmed in subsequent studies.

***Consider biologically relevant markers during signature development,* even if some are statistically weak on their own.**

Final Subset Derivation

Need subsets of markers that provide best predictive power.

Several options available. For example:

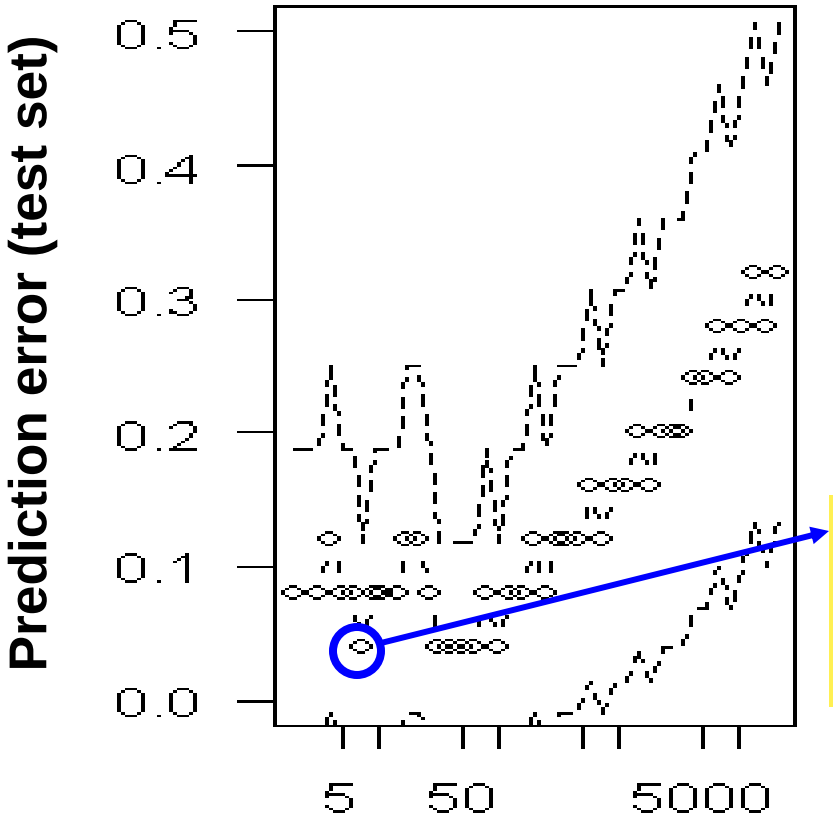
1. Using just the top few markers from the Filtering step.
2. Forward stepwise elimination (linear, logistic, GAM, etc.)
3. *Simulated Annealing with nearest centroids.*
4. *Genetic algorithm*
5. *Numerical optimization of relative importance scores from random forests or other models.*
6. *Shrunken Centroids & other model-fitting algorithms that have built-in subset selection options.*

- The subsets of markers derived from one of the above or other approaches can then be run in various classification algorithms.

Subset selection using Random Forests

Example: ~30K microarray experiment

Graphical representation of composite selection in RF using importance scores. Diaz-Uriarte (2006. BMC Bioinformatics)



Optimal Subset (7-genes) derived from a filtered list of ~5000 genes

Classification Algorithms

- Forward stepwise linear/logistic/GAM models
- Shrunken Centroids, Supervised Principal Components
- Random Forests, AdaBoost, Bagging, SVM, Neural Nets, SAM, KNN, Naïve-Bayes, etc.
- These and/or other algorithms/models are fit to the subset of markers derived in previous step.
- OK to mix & match the subset selection & model fitting methods.
  For example,
  - Fitting Neural Nets to a subset derived from Random Forests.
  - Fitting SVM to a subset derived from simulated-annealing.

Internal Validation

Using same data to identify and evaluate a biomarker signature will exaggerate the performance metrics.

Need Cross-Validation/Resampling, with several iterations.

<u>k-fold cross-validation:</u>

- Original data divided randomly into k equal parts
  - If N=100, k=5, obtain 5 random subsets of 20 each.
- Leave first part out, “train” on the remaining, “test” on the left-out.
- Repeat this for each of the other parts;
- Aggregate predictions from all left-out parts.
- Calculate performance metrics (e.g., sensitivity, specificity)
- Repeat this procedure 25 times. Report Mean & SD of the metrics.

Internal Validation (contd.)

Choice of k depends on N. Generally 5 to 10 is OK.

- Very small k can lead to fragile results & significant bias.

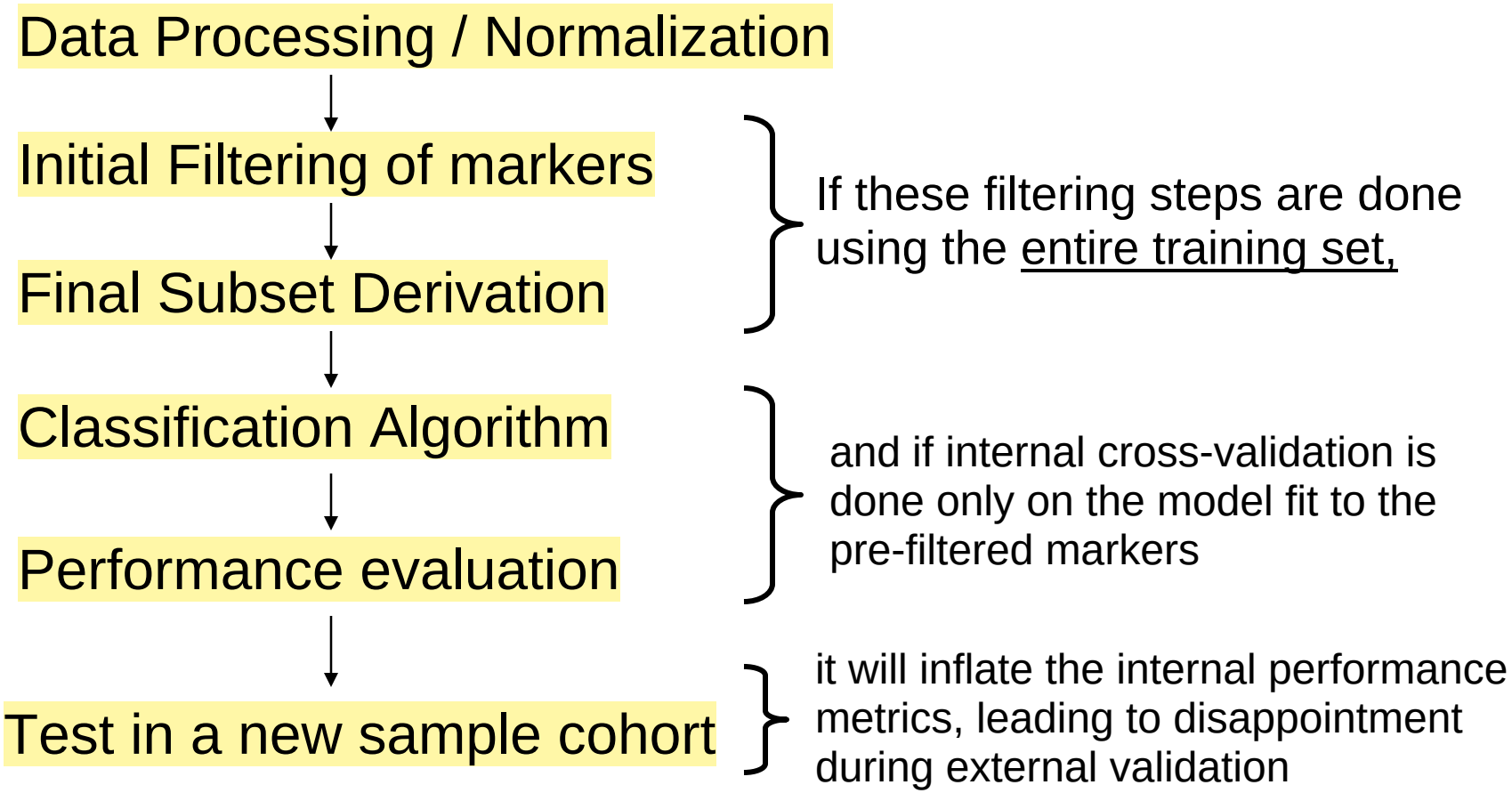
Example of Questionable results:

- Dave et al. *"Prediction of survival in follicular lymphoma based on molecular features of tumor infiltrating cells".* NEJM, Nov. 18, 2004 vol. 35set 2:2159-2169
- Reasons are explained and illustrated at:
  - http://www-stat.stanford.edu/~tibs/FL/report/index.html

→ *Don't take publication/literature claims for granted.*

Also, assess whether minor changes in the model can lead to major changes in the results…

Common error in performance evaluations



Fully embed all analysis steps within cross-validation

External Validation

After rigorous internal cross-validation, *test the signatures in independent external cohorts*.

- Should adequately represent the target population.

Samples in training & external sets are seldom run together.

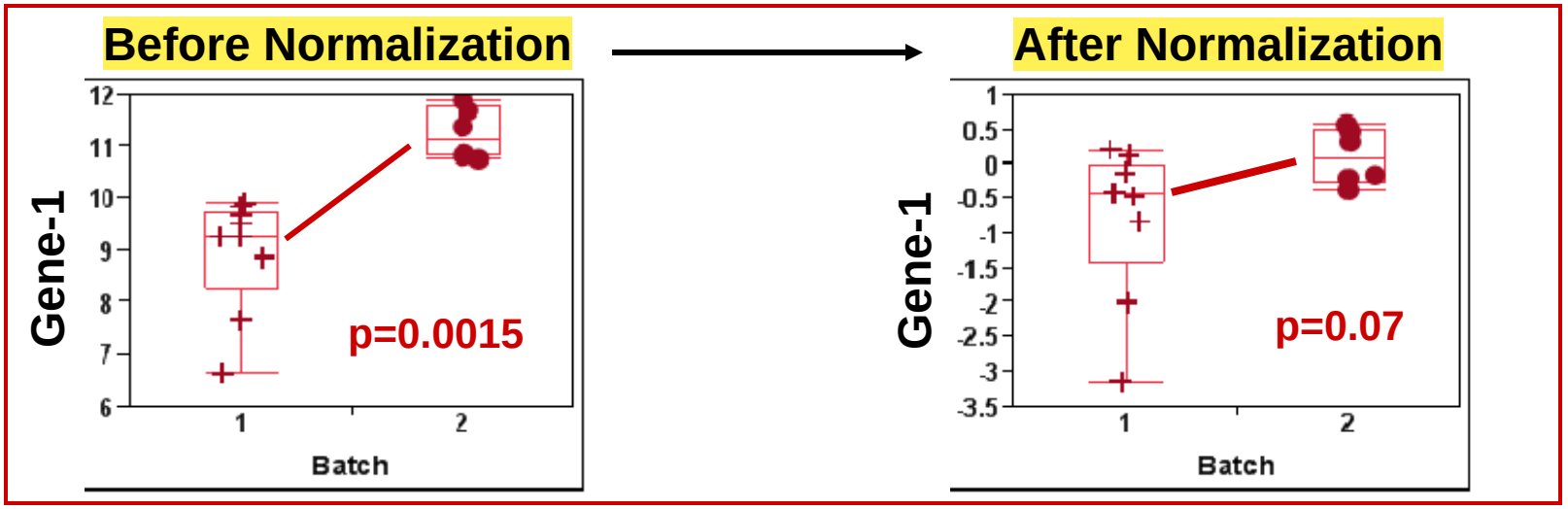
So *batch-effect normalization* may be necessary.

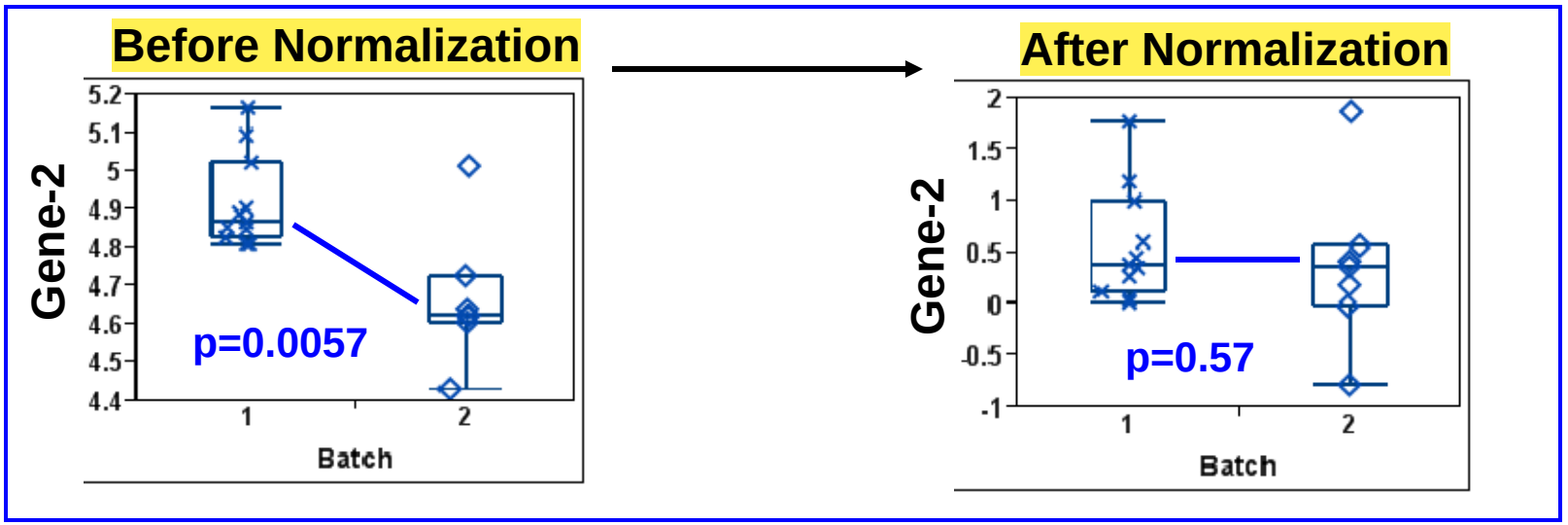
1. Normalize the training & external data.
   - A method that works well in my experience: *Eigen-Strat.*
2. Apply previously derived signature on the normalized training set.
3. Use this model on normalized external data to predict the response.

- *When applying signatures in future real samples, factors related to batch-effect (e.g., reference standard, reagents, etc) should be considered carefully.*

Batch-Effect in External Validation

Illustration from a genomics project

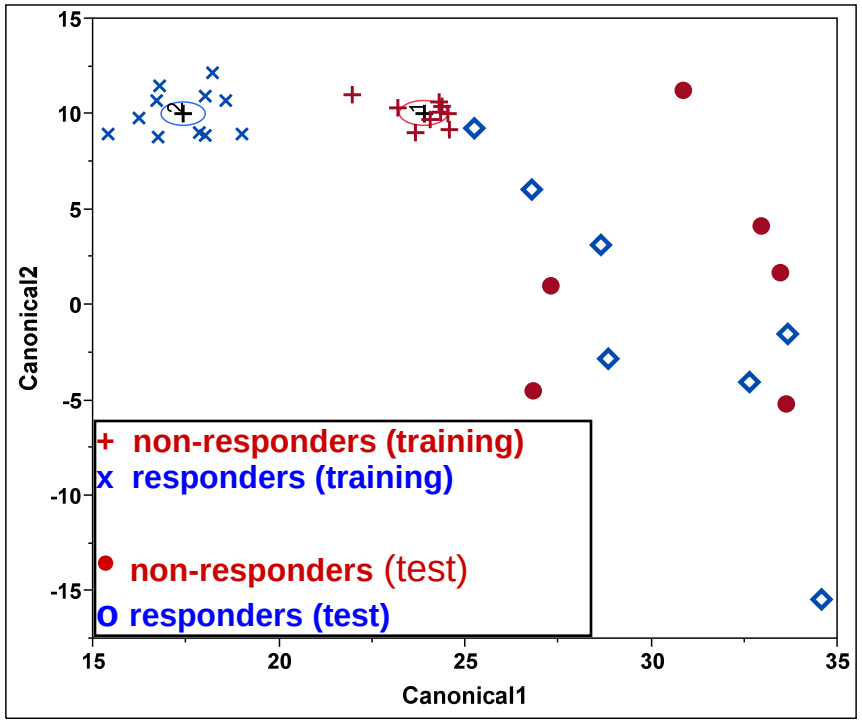




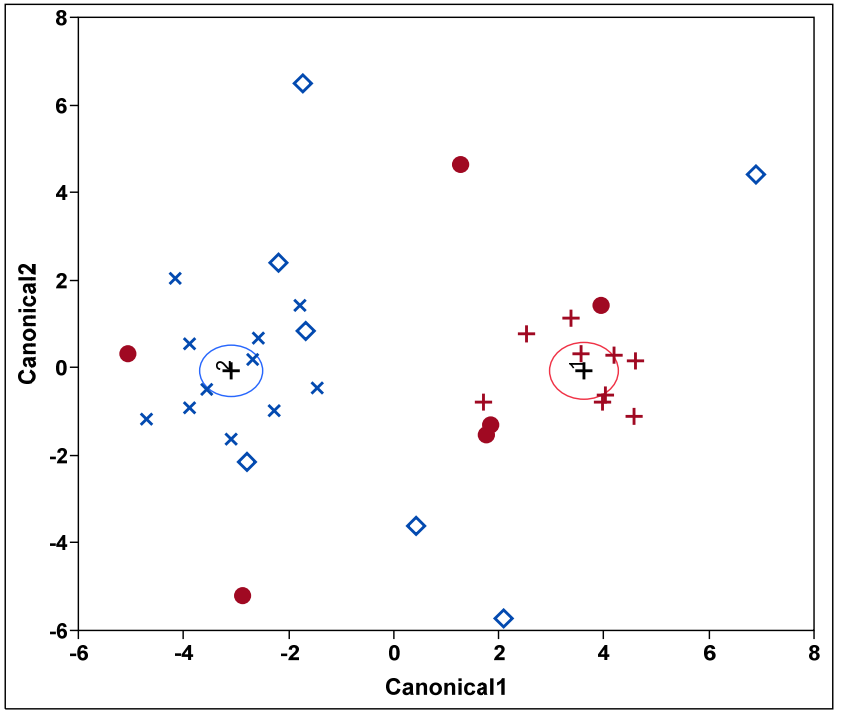
Impact of batch-effect on external validation

Illustration from a genomics project (contd.)

Before Normalization



After Normalization



Before normalization, all "responders" are incorrectly predicted.

Normalization yields significant improvement, although far from perfect due to other challenges (external set included a different disease state as well).

Exclusive focus on biological pathway markers

Instead of using whole array data, a biologically targeted list alone (e.g., from pathway analysis) can be used in the signature development process.

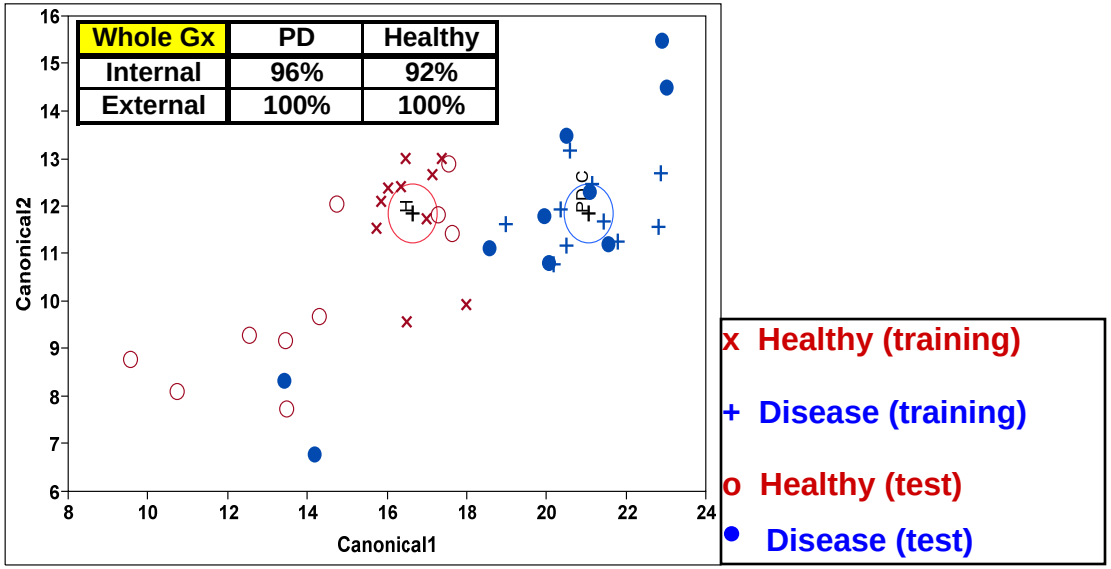
- i.e., instead of using data from all 50,000 probesets, one can work with just a biologically related list of say 1000 probesets.
- If the performance of these signatures are comparable to those from the whole genome, these will be very valuable.
  - Even if they are not as good, but within reasonable performance expectations, there may be interest in such signatures.

Clinicians & biologists may be more comfortable with signatures based on primarily biologically relevant markers.

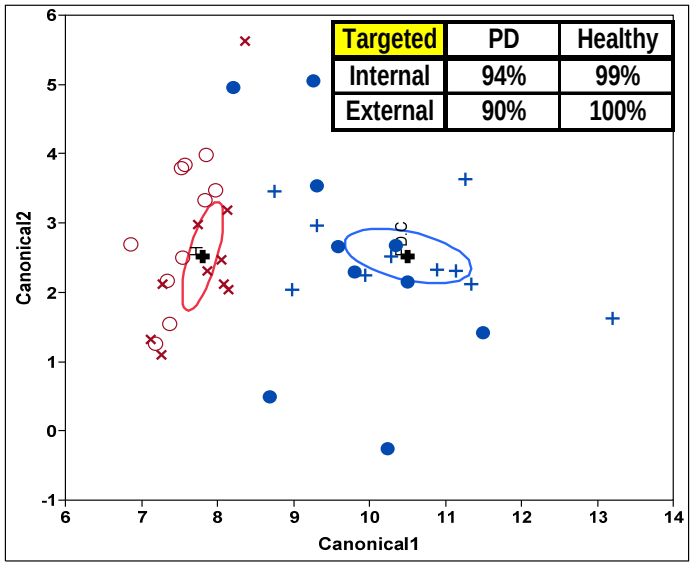
Exclusive focus on biological pathway markers

Case-Study

Signature from whole array



Signature from targeted list



Optimal signature from the smaller targeted list performs almost as well as the optimal signature from the whole array data.
Have seen this in at least two different disease applications.

This is not always possible or appropriate. But when the interest is primarily on disease/target-related markers, this is worth trying.

Summary

Biomarker Signature Development

Biomarkers come from a variety of sources. Often useful to consider markers across modalities for a signature.

Need statistical **and** biological criteria when filtering markers.

Variety of machine-learning methods needed during development. In the end, prefer a simpler model if performance is comparable to the more complex model.

Fully embedded cross-validation/resampling procedure needed to evaluate the performance.

Batch-effect factors (analysts, reagents, etc.) can impact external validation & real use of signatures. Normalize/control when possible.

Markers from disease pathways alone can yield similar performance but considering them in conjunction with novel markers often a good strategy.