Approaches to Evaluating and Validating Therapeutically Relevant Biomarkers

Annette Molinaro, Ph.D.
Division of Biostatistics
Yale University School of Medicine

What do we need?

- Improved tools for selecting individual patients for treatments
- Accurate prediction of who will respond and who will not.

What we have

- New technologies for genomic profiling
- Thus far, none have made it into clinical practice
- Prognostic factors will only be used if therapeutically relevant

Why?

- Clinical Drug Trial
  - Generally prospective
  - Patient Selection Criteria
  - Primary End Point
  - Stated hypotheses
  - Analysis plan specified in advance
  - Written protocol

- Prognostic Mkr Study
  - Frequently retrospective
  - No patient eligibility criteria
  - No primary end point
  - No stated hypotheses
  - No defined analysis plan
  - No written protocol

Consensus on Approach

- Developmental study
- Verify internal validity
- Translate to a common platform
- Verify reproducibility / external validity

Simon 2006. Ransohoff 2004. Barker 2003.

Maruvada et al 2006. Molinaro et al 2005.

What is a classifier?

- Mathematical function that maps the biomarker values to a set of prognostic categories (good risk, poor risk)

- Completely defined

What is validation?

“consists of efforts made to confirm the accuracy, precision, or effectiveness of results”

Feinstein, A.R. Multivariable Analysis: An Introduction (Yale University Press, New Haven, 1996)

What a classifier is not.

- A list of biomarkers or genes
  - – Correlated expression with outcome
    - Does not evaluate a defined diagnostic classifier which can be applied to patients
  - – Identified as associated with outcome
    - Unstable due to co-regulation within gene groups
    - Stringent criteria decreases statistical power

Such a list does not allow for prospective clinical validation

Developmental Study

- **Key:** To address a specific important therapeutic decision

- Analogous to Phase II of clinical trial
- Patients homogenous

- **Goal:** Completely specified classifier and corresponding hypotheses
  - Clinical value cannot be evaluated in the same study

Developing a Classifier

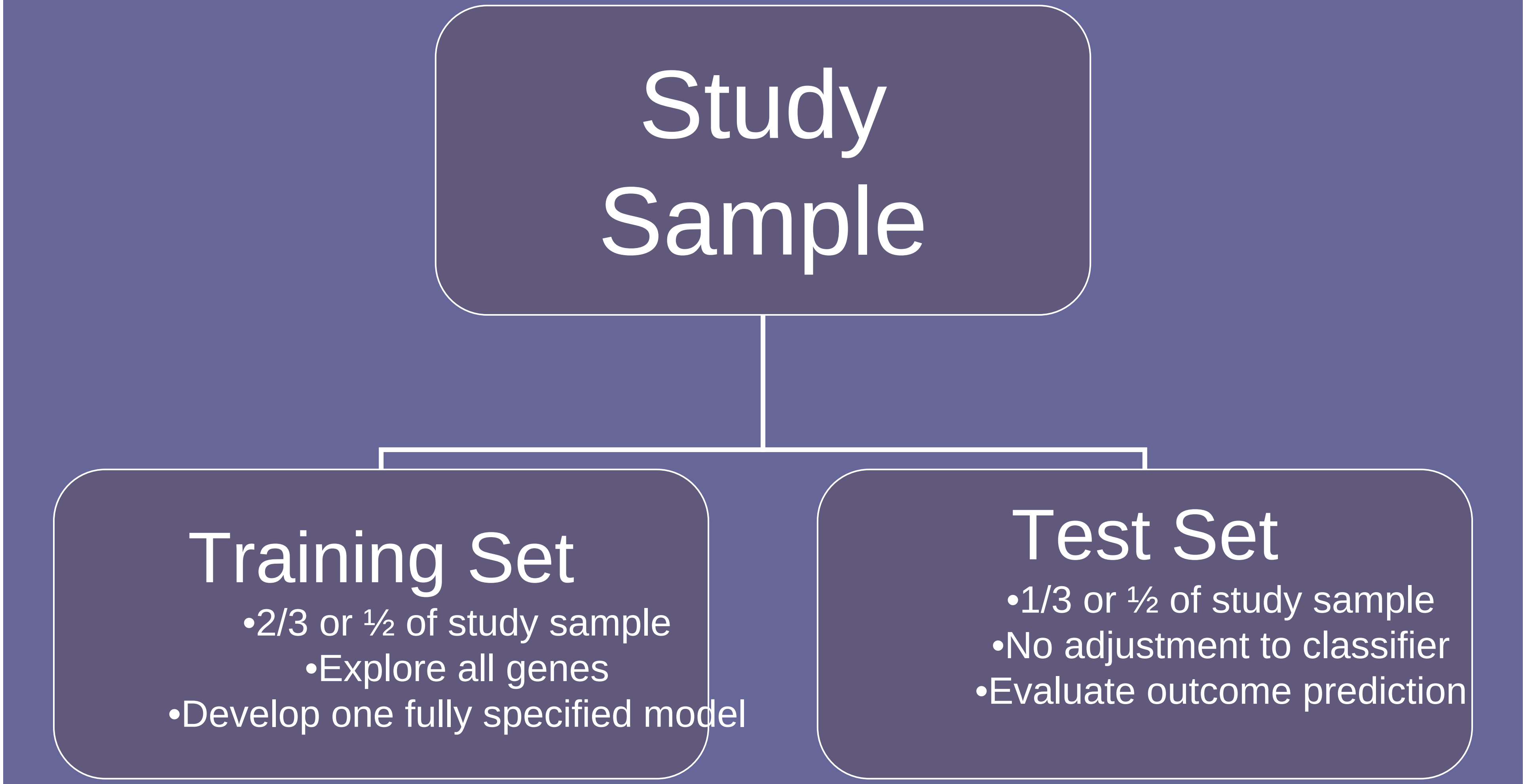
Main steps:

1. Prediction Model Selection
   - Many different algorithms
   - Number of genes much larger than number of observations
2. Split sample data into training & test set
3. Feature Selection
4. Fit model to training set
5. Estimate prediction accuracy with test set

Internal Validity

- Always possible to find perfect classifier even when no signal.
- To avoid ‘overfitting’ or ‘chance’ must use some form of training/test set
  - – Split Sample
  - – Cross-validation
- Important notes
  - – No adjustment of model or fitting on test set
  - – Feature selection is done within training set
- Assess statistical significance
  - – Estimate of prediction error
  - – Does the prediction error CI include chance?

Split Sample



Study Sample

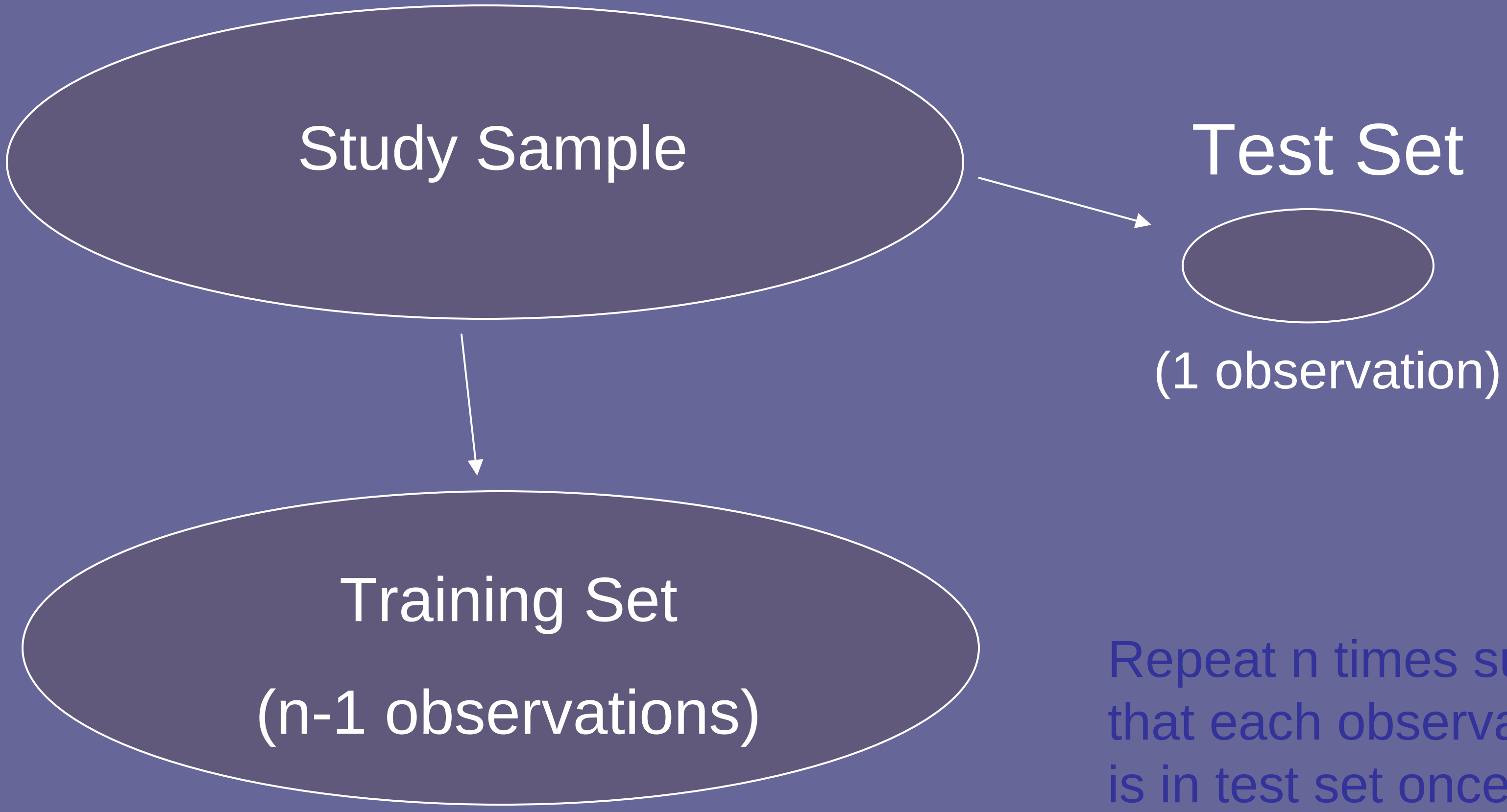
Training Set

- 2/3 or ½ of study sample
- Explore all genes
- Develop one fully specified model

Test Set

- 1/3 or ½ of study sample
- No adjustment to classifier
- Evaluate outcome prediction

Leave-One-Out Cross-Validation



Study Sample

Test Set

(1 observation)

Training Set

(n-1 observations)

Repeat n times such that each observation is in test set once.

Internal Validity

Estimate of prediction error for entire developmental study sample

<u>Questions answered:</u>

- Is classifier sufficiently accurate?
- Does it exceed or enhance the prediction accuracy of standard prognostic factors?
- Is it worthy of further investigation?

Example

BCCA-Herceptin Cohort

- 152 patients with metastatic breast cancer treated with Herceptin (trastuzumab) +/- concurrent systemic chemotherapy
  - » 61.4% taxol
  - » 22.9% vinorelbine

Giltnane, et al. In Preparation

Why did 52 not respond to treatment?

Table 1:
A) Univariate Logistic Models (Controlling for Concurrent Treatment)

		95% Confidence Intervals		
Variables	Odds Ratio	Lower	Upper	p-value
ER	**1.040**	**1.005**	**1.077**	**0.027**
PR	0.993	0.975	1.012	0.487
EGFR	0.996	0.982	1.010	0.571
HER2	**0.985**	**0.972**	**0.998**	**0.024**
HER3	1.012	0.996	1.028	0.153
HER4tm	1.012	0.984	1.040	0.409
HER4nuc	1.014	0.986	1.043	0.332
HER4mem	1.004	0.982	1.027	0.712

Giltnane, et al. In Preparation

Focus on predictive accuracy not on p-value

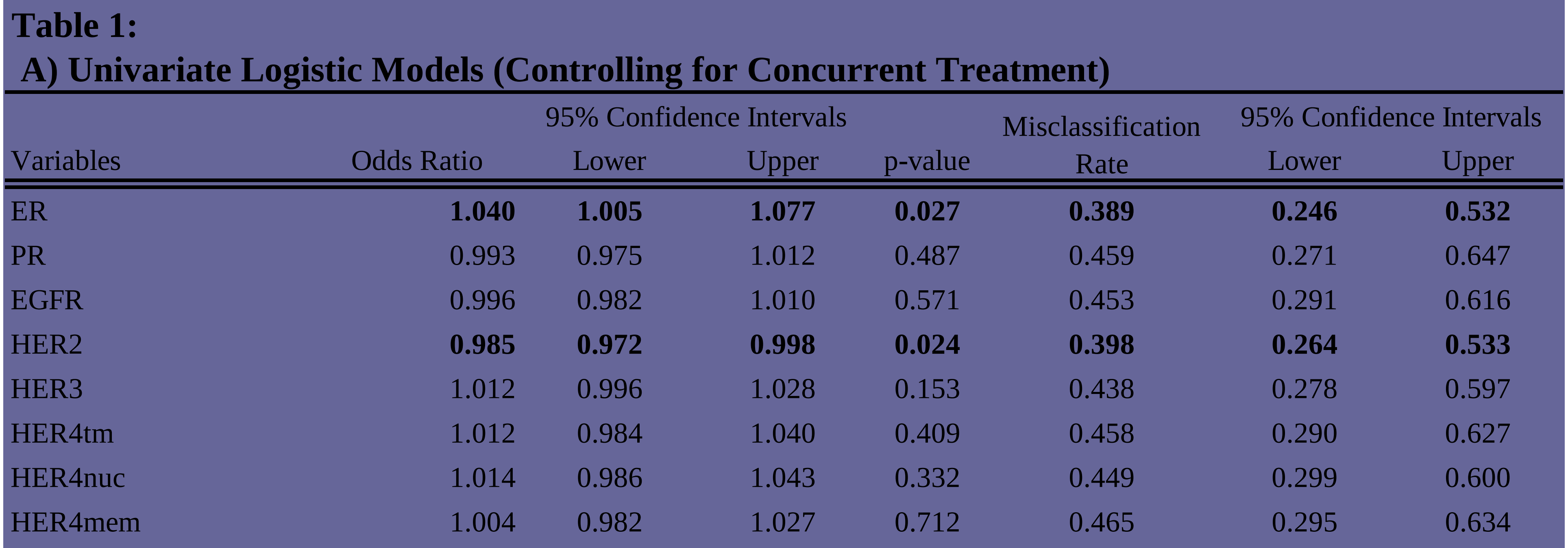


Table 1:

A) Univariate Logistic Models (Controlling for Concurrent Treatment)

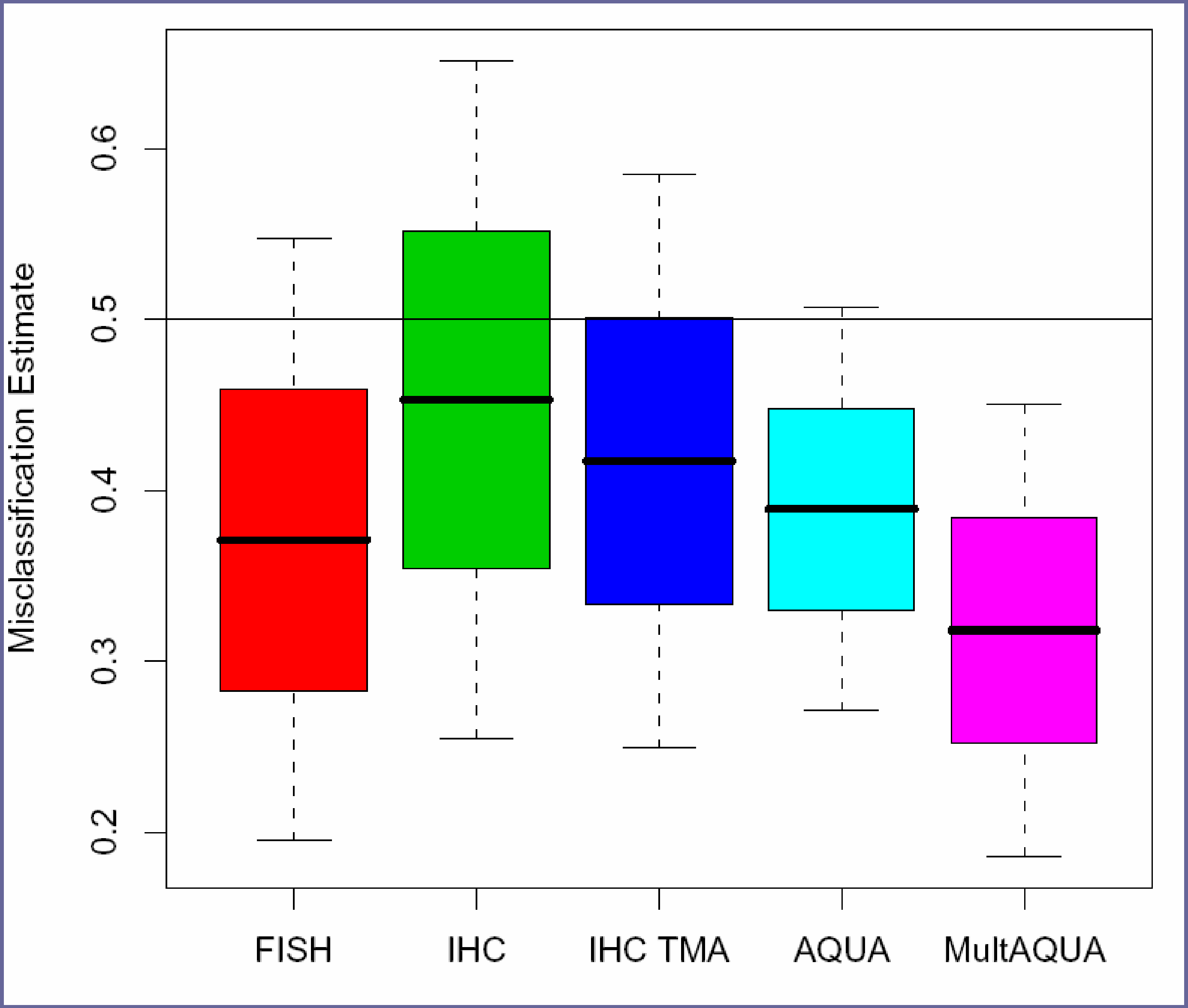
		95% Confidence Intervals			Misclassification	95% Confidence Intervals	
Variables	Odds Ratio	Lower	Upper	p-value	Rate	Lower	Upper
ER	**1.040**	**1.005**	**1.077**	**0.027**	**0.389**	**0.246**	**0.532**
PR	0.993	0.975	1.012	0.487	0.459	0.271	0.647
EGFR	0.996	0.982	1.010	0.571	0.453	0.291	0.616
HER2	**0.985**	**0.972**	**0.998**	**0.024**	**0.398**	**0.264**	**0.533**
HER3	1.012	0.996	1.028	0.153	0.438	0.278	0.597
HER4tm	1.012	0.984	1.040	0.409	0.458	0.290	0.627
HER4nuc	1.014	0.986	1.043	0.332	0.449	0.299	0.600
HER4mem	1.004	0.982	1.027	0.712	0.465	0.295	0.634

Giltnane, et al. In Preparation

B) Multivariate Logistic Model

		95% Confidence Intervals		
Variables	OddsRatio	Lower	Upper	p-value
ER	1.251	1.016	1.541	0.035
HER2	0.978	0.96	0.996	0.017
EFGR	1.031	1.002	1.06	0.033
ER*EGFR	0.996	0.992	0.999	0.024
HER4tm	1.318	1.012	1.718	0.041
HER4mem	0.836	0.705	0.992	0.04
HER4nuc	0.94	0.852	1.037	0.216
Rx-Taxol2	0.266	0.034	2.1	0.209
Rx-Vinorelbine3	0.104	0.015	0.711	0.021
Rx-Other4	0.257	0.031	2.132	0.208

		95% Confidence Intervals	
	Misclassification Rate	Lower	Upper
Model	**0.318**	**0.189**	**0.447**



Giltnane, et al. In Preparation

External Validation

Independent validation of prediction accuracy for completely specified classifier

- Prospective clinical trial
- Archived tissue

Determine if patient benefit (e.g. better efficacy, reduced incidence of adverse events, better convenience, lower costs) vs. not using the classifier.

Conclusions

- Assess prediction accuracy.
- Do not validate a classifier with the same data with which it was built.
- As editors, reviewers, and investigators verify internal and external validity.
- *“If overfitting issues have not been addressed then results should be regarded as inconclusive.”* (Ransohoff, 2004)

Acknowledgements

Yale University

- David Rimm
- Robert Camp
- Jena Giltnane

BCCA

- David Huntsman
- Karen Gelmon

NCI

- Richard Simon
- Ruth Pfeiffer

Funding Sources

- NCI K22 Career Transition Award (KCA123146A)

References

- Giltnane JM, Molinaro AM, Moeder C, Robinson A, Turbin D, Gelmon K, Huntsman DG, Rimm DL. Models Using Quantitative Multiplexed Assessment of Protein Expression to Predict Outcome in Breast Cancer. *In Submission*.
- Molinaro AM, Simon R, and Pfeiffer RM. Prediction Error Estimation: A Comparison of Resampling Methods. *Bioinformatics* 21(15):3301-3307, 2005.
- Simon R. Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. JCO 23(29):7332 – 7341, 2006.
- Barker PE. Cancer Biomarker Validation: Standards and Process Roles for the NIST. Ann. N.Y. Acad.Sci. 983: 142-150, 2003.
- Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nature Reviews Cancer 4: 309-314, 2004.
- Maruvada P and Srivastava S. Joint NCI-FDA Workshop on Resarch Strategies, Study Designs, and Statistical Approaches to Biomarker Validation for Cancer Diagnosis and Detection. Cancer Epi Bio Prev 15(6):1078-1082, 2006.
- Kattan MW. Judging new markers by their ability to improve predictive accuracy. JNCI 95(9):634 – 635, 2003.