Reporting and Evaluation of Studies of Biomarkers and Omics-based Predictors: *REMARK Guidelines and NCI Omics Checklist*

Canadian Statistical Sciences Institute (CANSSI) Workshop

Lisa McShane, PhD

Biometric Research Branch, DCTD

U.S. National Cancer Institute

November 7, 2014

Outline

- ❑ Background & definitions for tumor marker prognostic studies
- ❑ Role of reporting guidelines: REMARK
- ❑ Scaling up to omics-based predictors: cautions in study design and conduct
- ❑ Criteria to judge readiness of omics-based test to be used in a clinical trial
- ❑ Summary remarks

Definitions

- ❑ **Biomarker**

  http://www.cancer.gov/dictionary: "Biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease."

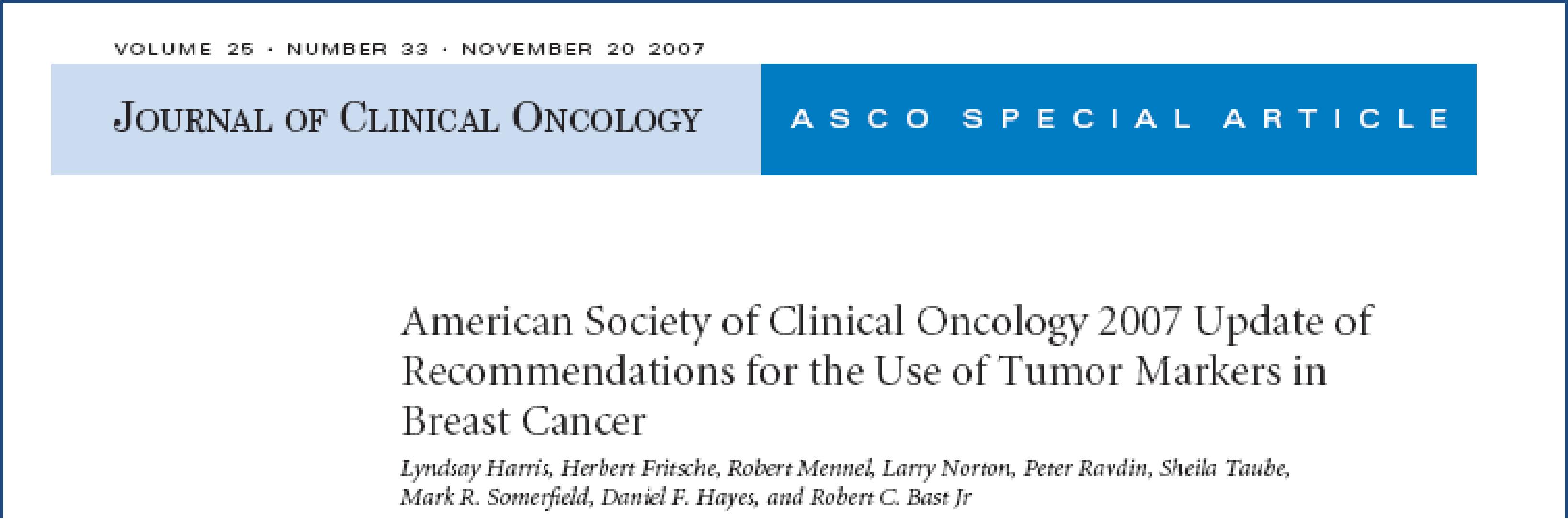
- ❑ **Prognostic**

  Associated with clinical outcome in absence of therapy (natural course) *or with standard therapy all patients are likely to receive*

  - May or may not be relevant for therapy decisions

FOCUS: Tumor prognostic markers

State of the Tumor Marker Literature



VOLUME 25 · NUMBER 33 · NOVEMBER 20 2007

JOURNAL OF CLINICAL ONCOLOGY

ASCO SPECIAL ARTICLE

American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer

Lyndsay Harris, Herbert Fritsche, Robert Mennel, Larry Norton, Peter Ravdin, Sheila Taube, Mark R. Somerfield, Daniel F. Hayes, and Robert C. Bast Jr

Purpose: To update the recommendations for the use of tumor marker tests in the prevention, screening, treatment, and surveillance of breast cancer.

“. . . primary literature is characterized by studies that included small patient numbers, that are retrospective, and that commonly perform multiple analyses until one reveals a statistically significant result. . .many tumor marker studies fail to include how patients were treated or analyses of the marker in different treatment subgroups. The Update Committee hopes that adherence to . . . REMARK criteria will provide more informative data sets in the future.

State of the Tumor Marker Literature

"Studies of 'prognostic' markers of no real future clinical utility and single biomarker studies will not be considered. Reports of studies into prognostic markers should be prospective and have a clear view of the practical clinical applications of the results. Retrospective analysis of biomarkers can be considered, if done within the framework of data collected from a prospective trial, with appropriate statistics and with multivariate analysis that includes established predictive/prognostic markers. Reports of prognostic tumor marker studies should follow the REMARK guidelines (available from www.equator-network.org)."

J. B. Vermorken

Editor-in-Chief

Statement of editorial intent
Annals of Oncology 2012; 23:1931-1932

REMARK: **RE**porting guidelines for tumor **MARK**er prognostic studies

Lisa M. McShane, Douglas G. Altman, Willi Sauerbrei, Sheila E. Taube, Massimo Gion, and Gary M. Clark for the Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics (*J Natl Cancer Inst* 2005; 97:1180-1184, and simultaneously in *BJC, EJC, JCO, NCPO)*

Recommended reporting elements to facilitate

- ❑ Evaluation of **appropriateness & quality** of study design, methods, and analysis
- ❑ Understanding of **context** in which conclusions apply
- ❑ **Reproducibility**
- ❑ **Comparisons** across studies, including formal meta-analyses

REMARK: Target Studies

- ❑ Studies relating marker values to clinical events (e.g., recurrence, death, response)
- ❑ NOT primarily aimed at biological discovery studies, but use encouraged to extent possible
  - ▪ Patients
  - ▪ Specimens
  - ▪ Assays
- ❑ NOT sufficient for studies developing multiplex classifiers/risk scores (e.g., derived from omics data), but applicable to studies assessing them

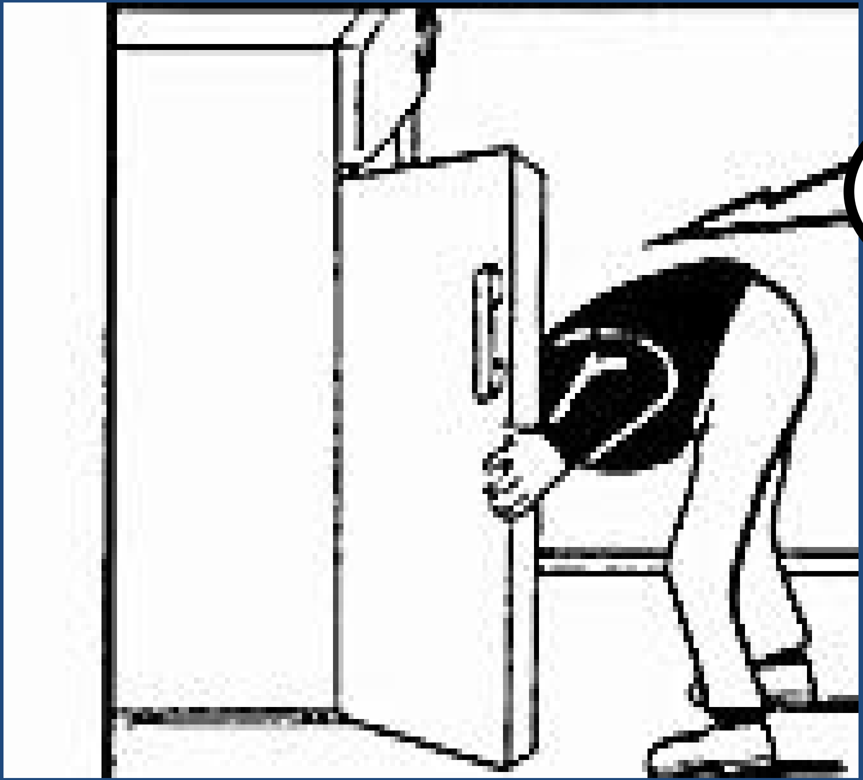
REMARK Elements: Introduction

❑ State all marker(s) examined

❑ Study objectives

❑ Pre-specified hypotheses

Common Tumor Marker Study Design



What can we do with our marker on these 89 specimens?

- “Convenience” specimens
- Heterogeneous patient characteristics
- Treatments: Unknown, non-randomized, not standardized
- Insufficient sample size (underpowered)
- Uncertain specimen and data quality

REMARK Elements: Materials & Methods

❑ Patients

- Inclusion/exclusion (e.g., stage, subtype), source, treatments

❑ Specimen characteristics

- Format, collection, preservation, storage
- See BRISQ criteria (Moore et al, *Cancer Cytopathology* 2011; 119:92-101)

REMARK Elements: Materials & Methods (cont.)

- Assay methods
  - Detailed protocol (reagents/kits), quantitation, scoring & reporting, reproducibility, blinding

Example: Systematic review (43 studies) of Ki67 in early breast cancer (Stuart-Harris et al, *The Breast* 2008; 17:323-334)

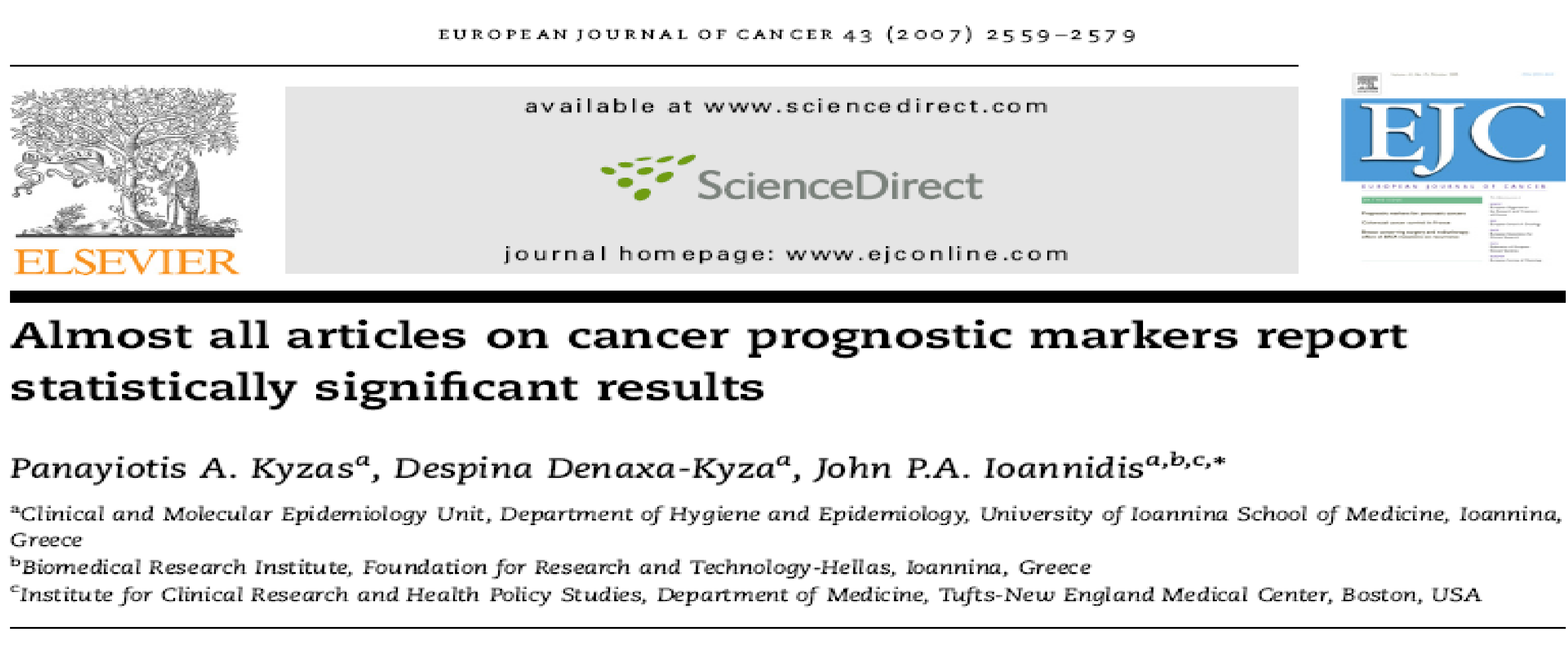
- English publication, Jan. 1995 – Sept. 2004
- ≥ 100 patients, OS or DFS endpoint

- Results
  - 7 different antibodies for IHC, single or combination
  - 19 different cutpoints, ranging from 0-30%
  - Significant between-study heterogeneity and evidence for publication bias

REMARK Elements: Materials & Methods (cont.)

- Study design
  - Case selection (e.g., random, case-control), clinical endpoints, variables considered, sample size
- Statistical analysis methods
  - Models, variable selection, handling of missing data, multiple testing adjustments, validations

Importance of identifying exploratory statistical analyses



EUROPEAN JOURNAL OF CANCER 43 (2007) 2559–2579

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Almost all articles on cancer prognostic markers report statistically significant results

Panayiotis A. Kyzas[a], Despina Denaxa-Kyza[a], John P.A. Ioannidis[a,b,c,*]

[a]Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece
[b]Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece
[c]Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Boston, USA

“If you torture the data long enough they will confess to anything.”

Source unknown

Statistical Analysis: Multiple Testing

- Multiple markers
- Multiple endpoints
- Multiple subgroups
- Multiple marker cutpoints
- Multiple models with multiple variables

Example: 8 subgroups defined by 3 binary factors

Number of independent tests ($\alpha = 0.05$ per test)	Probability observe ≥ 1 statistically significant ($p<0.05$) result
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
6	0.26
7	0.30
8	0.34
9	0.37
10	0.40

Statistical Analysis: Cutpoint optimization

- ❑ 203 patients with lymph node negative primary breast cancer
- ❑ Proliferation marker Ki67 measured by IHC on 193/203
- ❑ No adjuvant systemic therapy (chemo or endocrine)

Endpoint = Breast Cancer Specific Survival (BCSS)

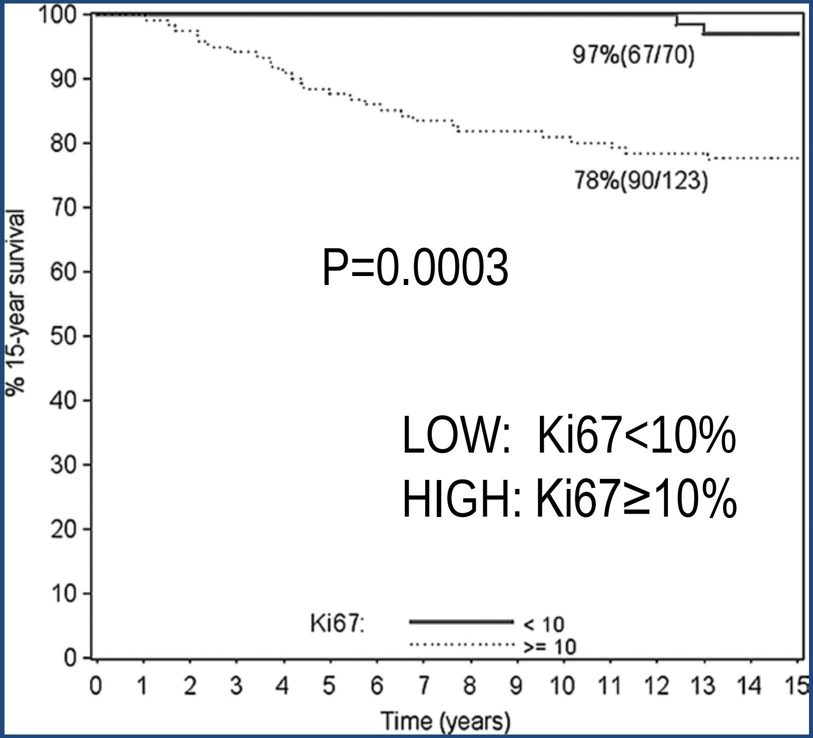
<u>LOW: Ki67<10%</u>

15-yr BCSS = 97%

<u>HIGH: Ki67≥10%</u>

15-yr BCSS = 78%

BCSS for low and high Ki67



Pathmanathan N et al. J Clin Pathol 2014;67:222-228

Statistical Analysis: Cutpoint optimization

Number of deaths, sensitivities and specificities according to a range of cut-off values of Ki67

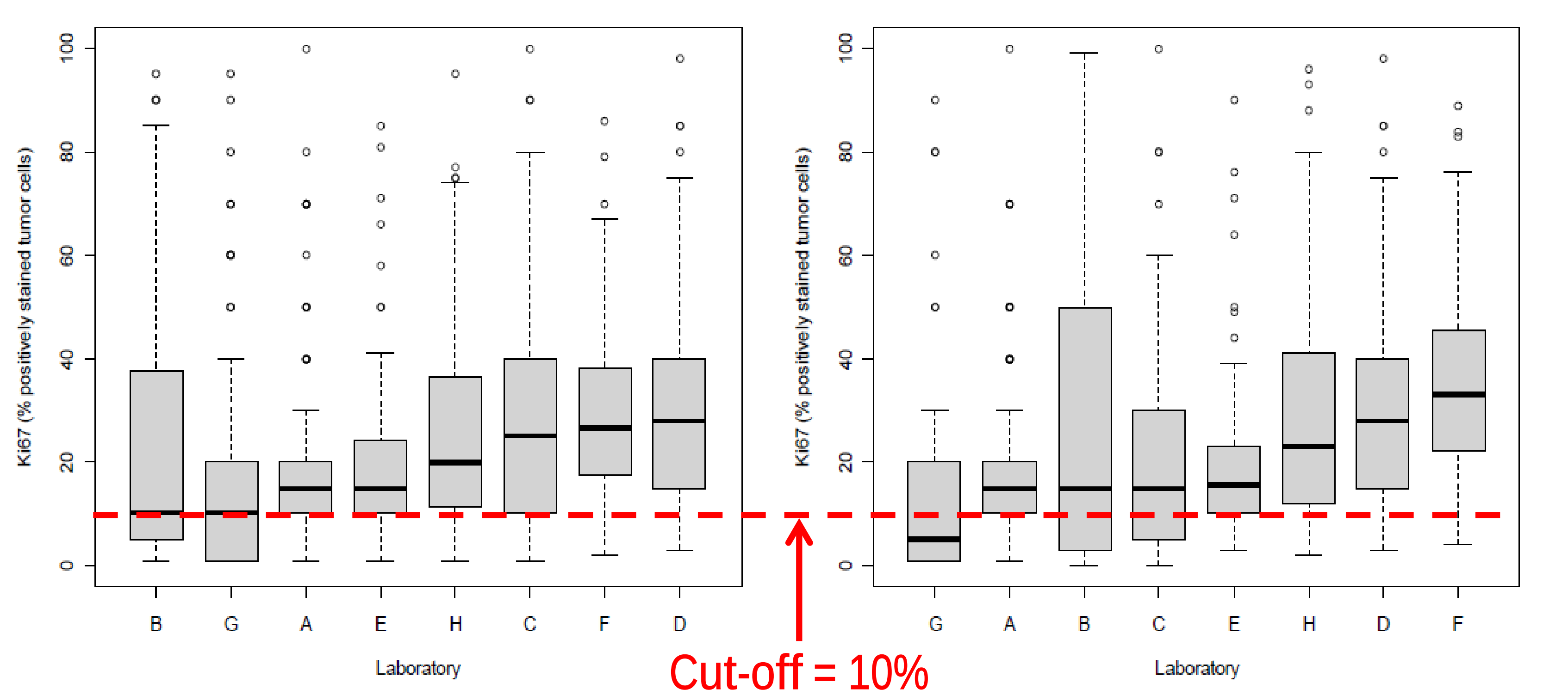
Ki67	No. died (%)	No. in category	Sensitivity	Specificity	Youden's index (J)
≥0	29 (15.0)	193	1	0	0
≥5	28 (17.6)	159	0.966	0.201	0.167
≥10	27 (22.0)	123	0.931	0.415	**0.346**
≥15	20 (21.3)	94	0.690	0.549	0.238
≥20	16 (22.5)	71	0.552	0.665	0.216
≥30	12 (25.0)	48	0.414	0.780	0.194
≥40	10 (27.8)	36	0.345	0.841	0.186
≥50	8 (27.6)	29	0.276	0.872	0.148

J = Sensitivity + Specificity – 1

Pathmanathan N et al. J Clin Pathol 2014;67:222-228 (Table 1)

Statistical Analysis: Cutpoint optimization and impact on assay transportability

Side-by-side boxplots of Ki67 distributions with 8 labs assessing different TMA sections of same set of 100 breast cancer cases



Cut-off = 10%

Centrally stained, locally scored
Median range: 10% to 28%

Locally stained, locally scored
Median range: 5% to 33%

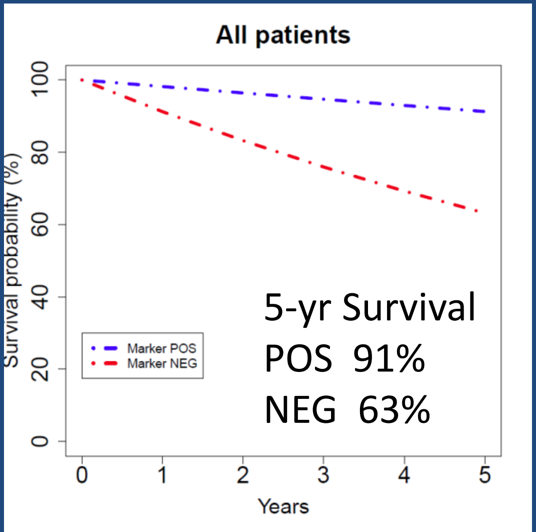
Polley M et al, J Natl Cancer Inst 2013; 105: 1897-1906 (Figure 2)

REMARK Elements: Results

- ❑ Data
  - Numbers of patients and events
  - Demographic characteristics
  - Standard prognostic variable distribution
  - Tumor marker distribution
- ❑ Analysis & presentation
  - Univariate analyses (marker vs. standard prognostic variables, marker vs. outcome)
  - Multivariable analyses (association of marker with outcome after adjustment for standard prognostic variables)
  - Measures of uncertainty for reported effect estimates

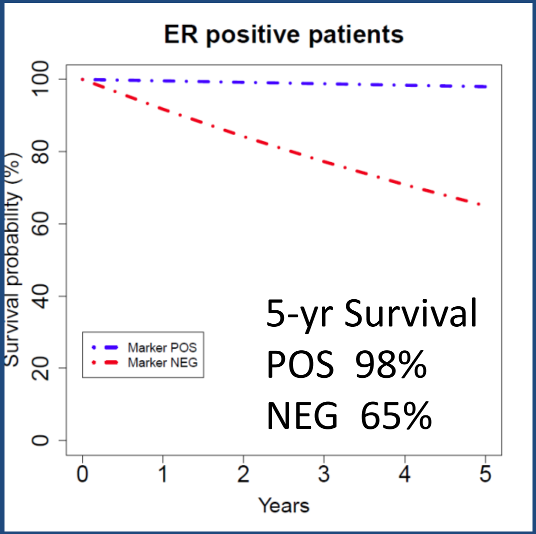
REMARK Elements: Results (cont.)

- Multivariate analysis vs. subgroups
  - Subgroup analyses may be important for interpretation
  - Better yet, study more clinically homogenous populations



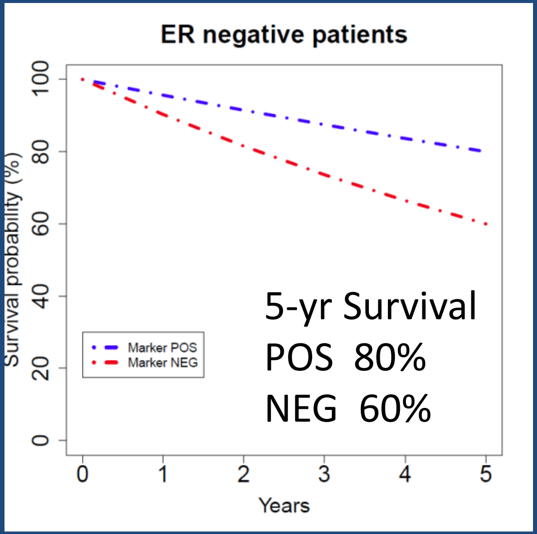
5-yr Survival
POS 91%
NEG 63%

65%



5-yr Survival
POS 98%
NEG 65%

35%



5-yr Survival
POS 80%
NEG 60%

REMARK Elements: Discussion

- ❑ Interpretation in context of pre-specified hypotheses
- ❑ Relevance to other studies
- ❑ Limitations
- ❑ Future research
- ❑ Clinical value

REMARK Status & Future

- □ Explanation & Elaboration: Altman et al, *PLoS Medicine* 2012; 9(5):e1001216 (also *BMC Medicine* 2012; 10:51)
- □ "Before vs. after" reporting quality
  - Before: Mallett et al, British Journal of Cancer 2010; 102: 173-180
  - After: Underway
- □ Journals stating REMARK adherence requirements: *Ann Oncol, Breast Cancer Res Treat, Clin Cancer Res, J Clin Oncol, J Natl Cancer Inst, J Pathol*

Scaling up to omics-based predictors

- ❑ Omics

  “A term encompassing multiple molecular disciplines, which involve the characterization of global sets of biological molecules such as DNAs, RNAs, proteins, and metabolites.”

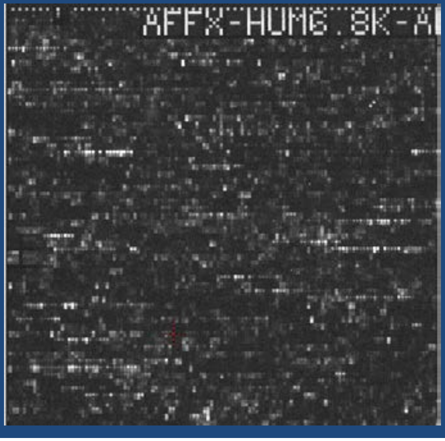
- ❑ Omics-based test

  “An assay composed of or derived from multiple molecular measurements and interpreted by a fully specified computational model to produce a clinically actionable result.”

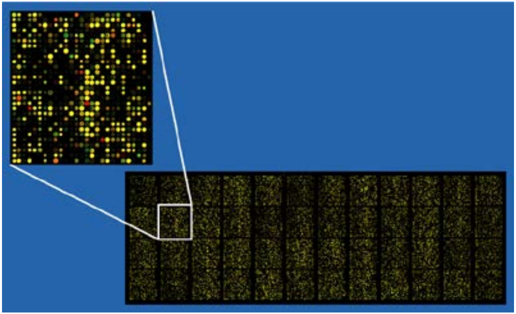
  (Mathematical model component referred to as a predictor or classifier with outputs such as risk score or categorization.)

Institute of Medicine report: Evolution of Translational Omics
http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx

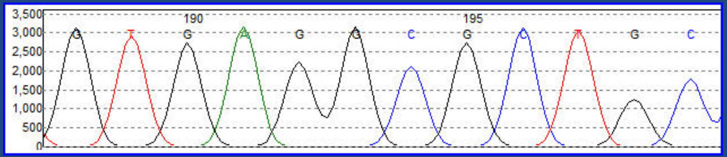
Omics assays



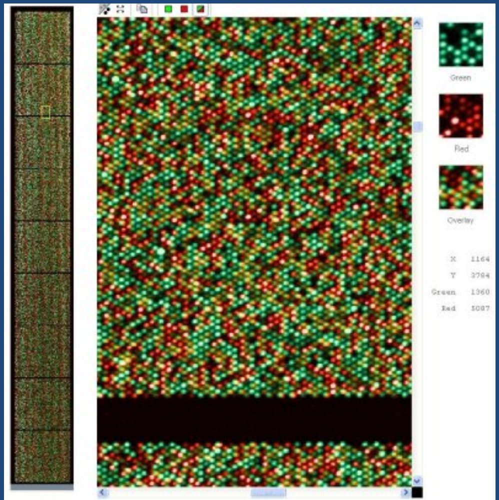
Affymetrix expression GeneChip



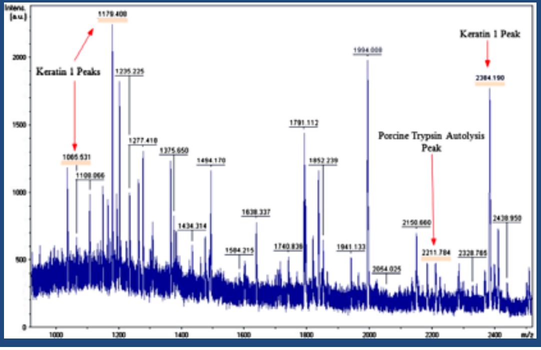
cDNA expression microarray



Mutation sequence surveyor trace



Illumina SNP bead array



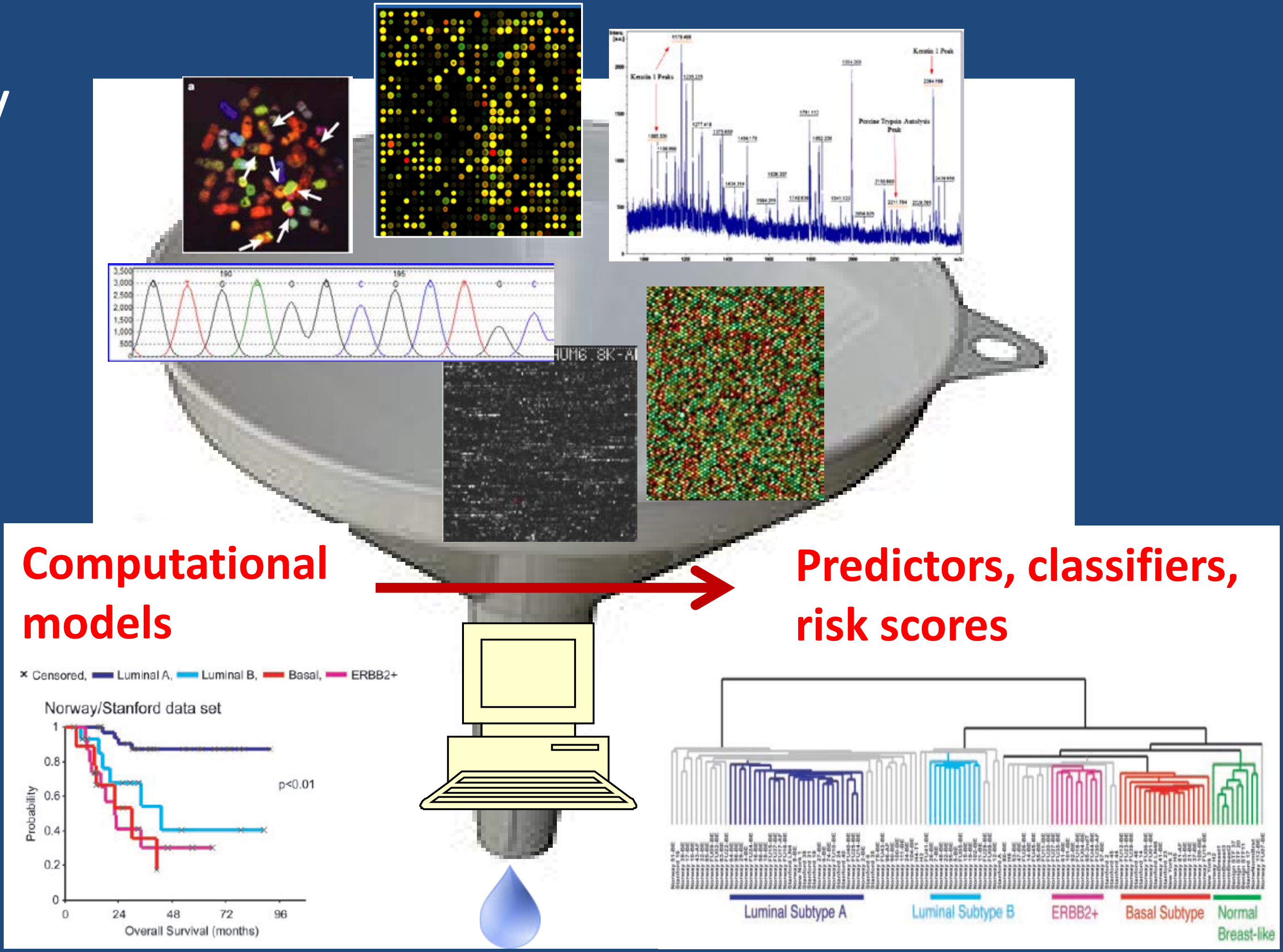
MALDI-TOF proteomic spectrum

Translation from omics discoveries to clinically useful omics-based tests

High-throughput omics assays

Discovery

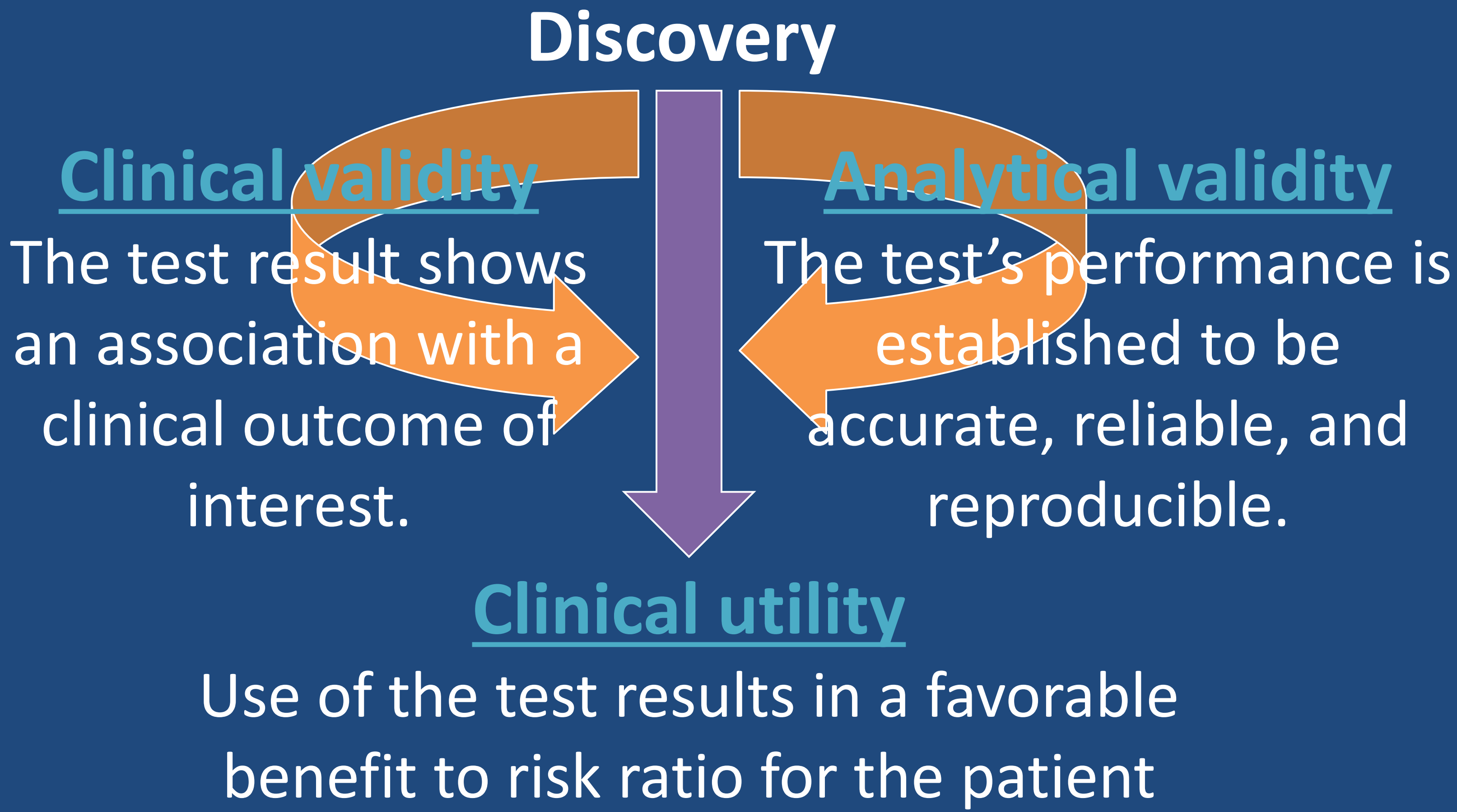
Clinical Utility?



Computational models

Predictors, classifiers, risk scores

Paradigm for development of a clinically useful omics-based test



Discovery

Clinical validity

The test result shows an association with a clinical outcome of interest.

Analytical validity

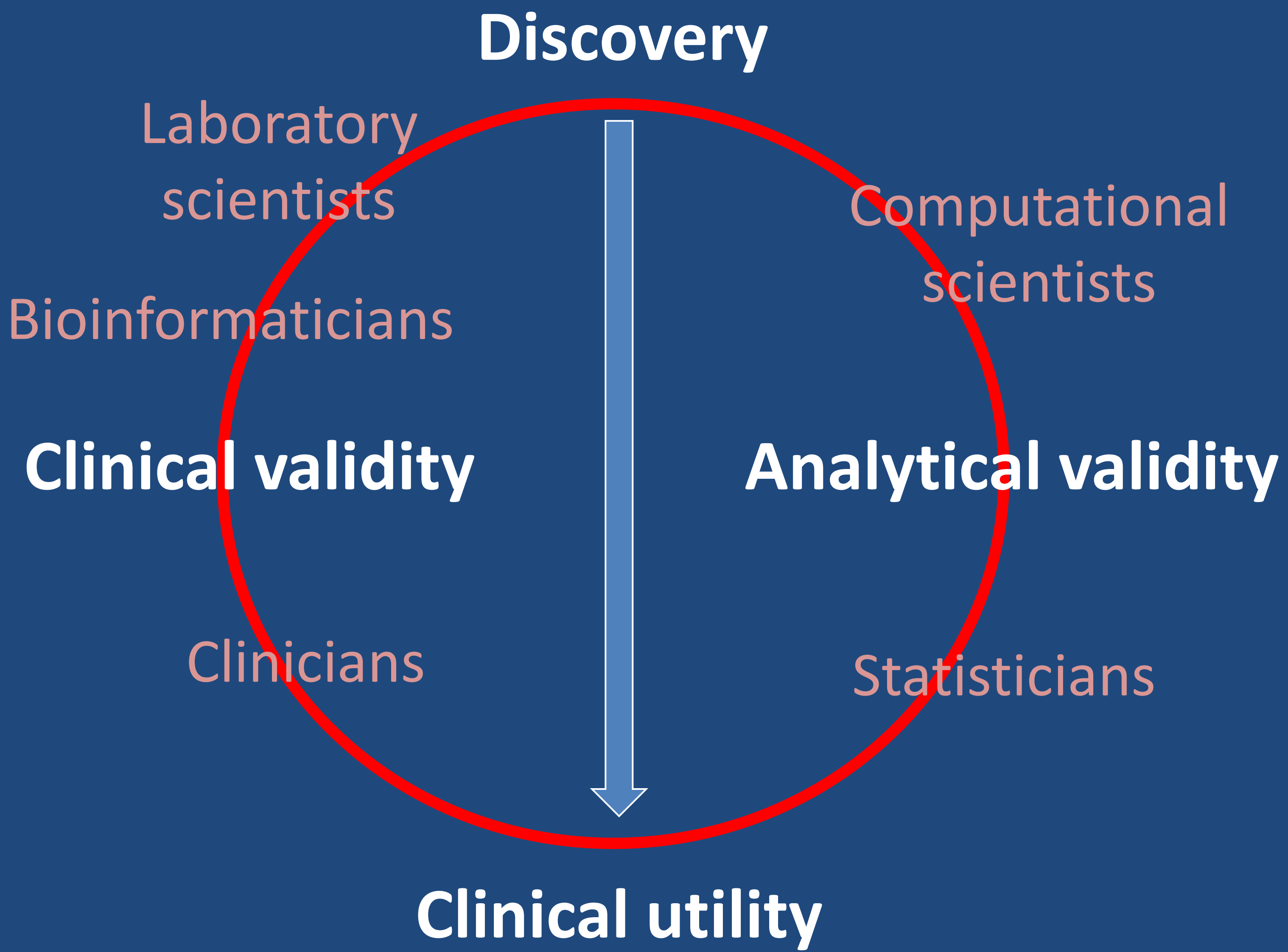
The test's performance is established to be accurate, reliable, and reproducible.

Clinical utility

Use of the test results in a favorable benefit to risk ratio for the patient

Teutsch et al, *Genet Med* 2009;11:3-14
Simon et al, *J Natl Cancer Inst* 2009;101:1446-1452
McShane & Hayes, *J Clin Oncol* 2012;30:4223-4232

It takes a collaborative team to go from discovery to clinically useful omics test



Discovery

Laboratory scientists

Computational scientists

Bioinformaticians

Clinical validity

Analytical validity

Clinicians

Statisticians

Clinical utility

NCI Criteria for the use of omics-based predictors in clinical trials

- Focus: Tests based on potentially complex mathematical models incorporating large numbers of measurements from omics assays
- Goals:
  - Make omics test development more efficient, reliable, and transparent
  - Avoid premature clinical implementation of omics-based tests

McShane et al, *Nature* 2013;502:317-320

McShane et al, *BMC Medicine 2013;*11:220

Omics checklist divided into 5 domains

- ❑ Specimens
- ❑ Assays
- ❑ Model development, specification & preliminary performance evaluation
- ❑ Clinical trial design
- ❑ Ethical, legal, and regulatory

Domain 1: Specimens

- ❑ Collection, processing & storage
- ❑ Specimen quality screening
- ❑ Minimum required amount
- ❑ Feasibility of collecting needed specimens
  - ▪ Achievable in standard clinical settings
  - ▪ Study/sample size planning

Domain 1: Specimens example

- ❑ Statisticians can provide guidance in planning feasibility assessments and quality monitoring schemes to avoid disasters

  Example:

  - ▪ Analysis of first 100 biological specimens collected in a large diagnostic study showed that only 20% were of adequate quality to be analyzable by the assay
  - ▪ Problem traced to failure to promptly freeze the specimens after collection

Domain 2: Assays

- ❑ Impact of changes in assay procedures
- ❑ Lock down SOP
- ❑ Quality criteria for assay values
  - ▪ Bad specimens, batch effects, equipment malfunction
- ❑ Analytical performance evaluation

  Pennello, *Clinical Trials* 2013;10: 666–676

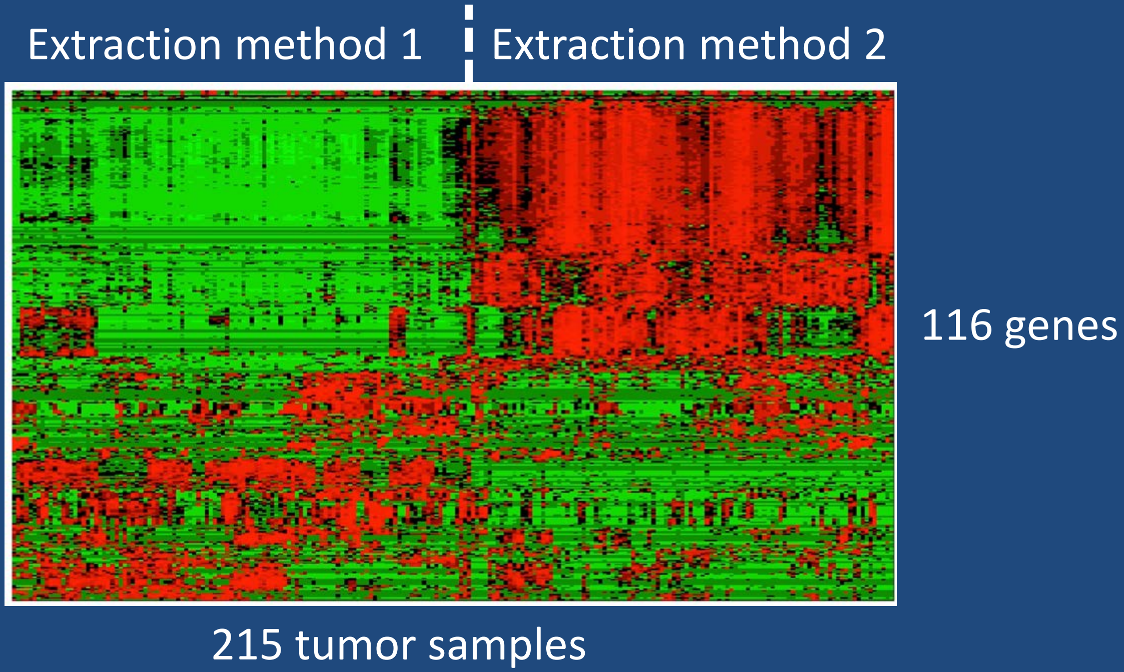
  Jennings et al, *Arch Pathol Lab Med* 2009;133: 743–755
- ❑ Quality monitoring
- ❑ Turnaround time

Domain 2: Assay example

- ❑ Assess impact of changes in any assay procedures, reagents, or equipment

  Example:

  Dramatic effect of change in RNA extraction procedure on tumor gene expression microarray profiles, additional minor effect due to reagent changes by microarray manufacturer



Domain 3: Model development & evaluation

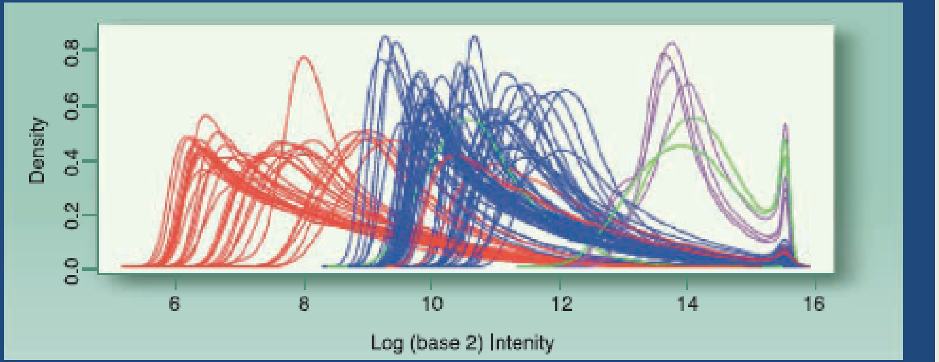
- ❑ Quality of data (clinical & omics) used to develop and validate predictor models
- ❑ Appropriate statistical approaches for model development and performance assessment
- ❑ Intended use - data from clinically relevant patient population

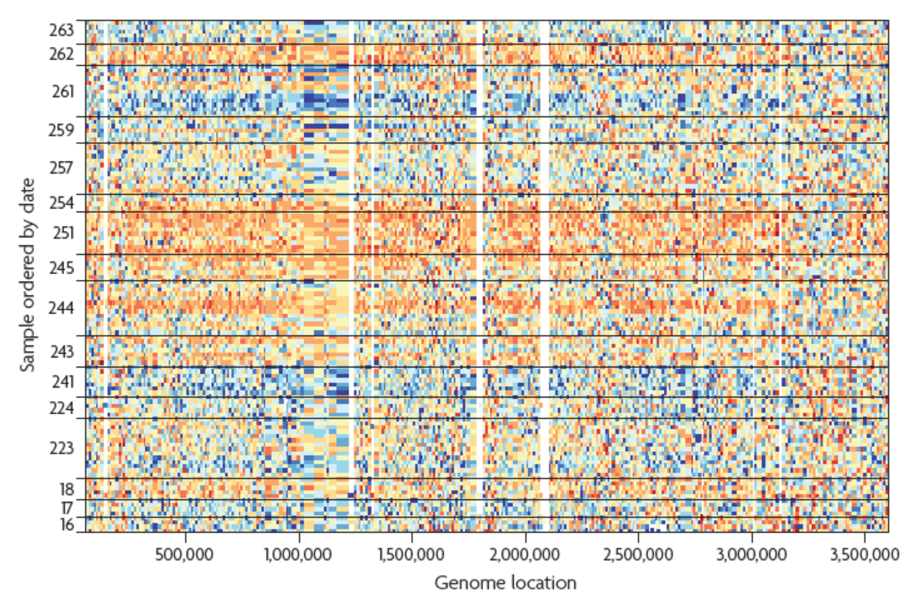
Domain 3: Data quality & batch effects

Density estimates of PM probe intensities (Affymetrix CEL files) for 96 NSCLC specimens

Red = batch 1
Blue = batch 2
Purple & Green = outliers?

(Owzar et al, *Clin Cancer Res* 2008;14:5959-5966)





Batch effects for 2nd generation sequence data (stand. coverage data).
Same facility & platform.
Horizontal lines divide by date.

(Leek et al, *Nature Rev Genet* 2010;11:733-739)

BATCH EFFECTS ARE ESPECIALLY PROBLEMATIC IF CONFOUNDED WITH KEY EXPERIMENTAL FACTORS OR ENDPOINTS.

Domain 3: Dangers of overfitting

- ❑ A statistical model is **OVERFIT** when it describes random error (noise) instead of the true underlying relationship
  - Excessively complex (too many parameters or predictor variables)
  - Generally has poor predictive performance on an independent data set

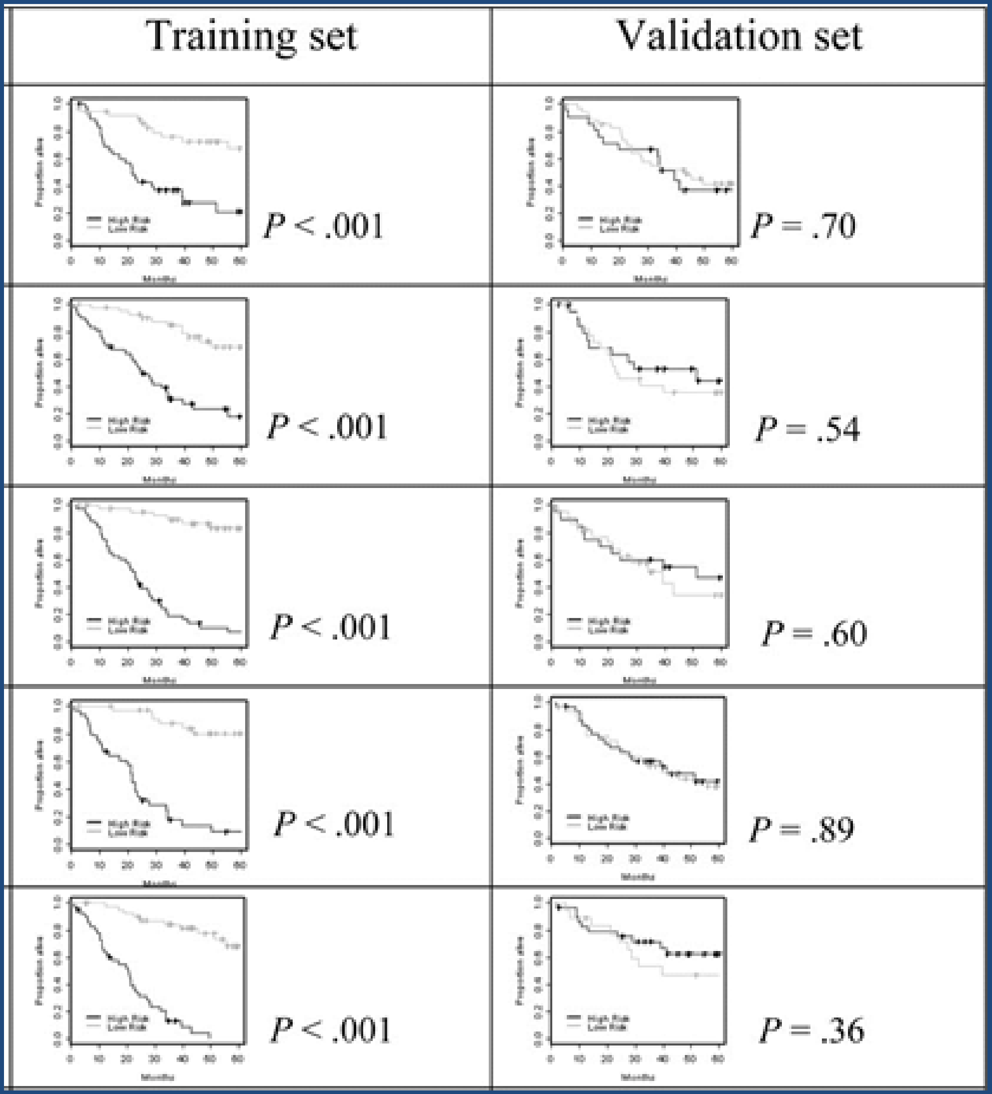
Domain 3: Failure to detect overfitting

- ❑ **RESUBSTITUTION** is the naïve practice of evaluating performance of a model by "plugging in" exact same data used to build it
  - ▪ Seriously biased estimates of predictor performance
  - ▪ Overfitting will not be detected

Domain 3: Avoid overfitting & resubstitution

Simulation of bias in resubstitution estimates of predictor performance

- Goal: Develop prognostic signature from gene expression microarray data
- Survival data on 129 lung cancer patients (prior study)
- Expression values for 5000 genes generated randomly from $N(0, I_{5000})$ (“noise”) for each patient
- Data divided randomly into training and validation sets
- Prognostic model developed from training set and used to classify patients in both training and validation sets (supervised principal components method)



(Subramanian & Simon, *J Natl Cancer Inst* 2010;102:464-474)

Domain 3: Detection and avoidance of model overfitting

- ❑ Internal validation by use of data resampling techniques
  - ▪ Split sample (training & test sets)
  - ▪ Cross-validation
  - ▪ Bootstrapping

  Molinaro et al, *Bioinformatics* 2005;21:3301-3307
- ❑ External validation
  - ▪ Assessment of predictor performance on a completely independent data set
- ❑ Model regularization techniques reduce, but don’t completely eliminate overfitting

Domain 3: Subtle forms of model overfitting

- ❑ Partial resubstitution
- ❑ Combining training and test sets
- ❑ Resubstitution with covariate adjustment
- ❑ Resubstitution comparison

Simon et al, *J Natl Cancer Inst* 2003;95:14-18
Subramanian & Simon, *J Natl Cancer Inst* 2010;102:464-474
Simon & Freidlin, [Correspondence] J Natl Cancer Inst 2012;103(5):445
Subramanian & Simon, *Contemporary Clinical Trials* 2013;36:636–641
McShane & Polley, *Clinical Trials* 2013;10:653-665

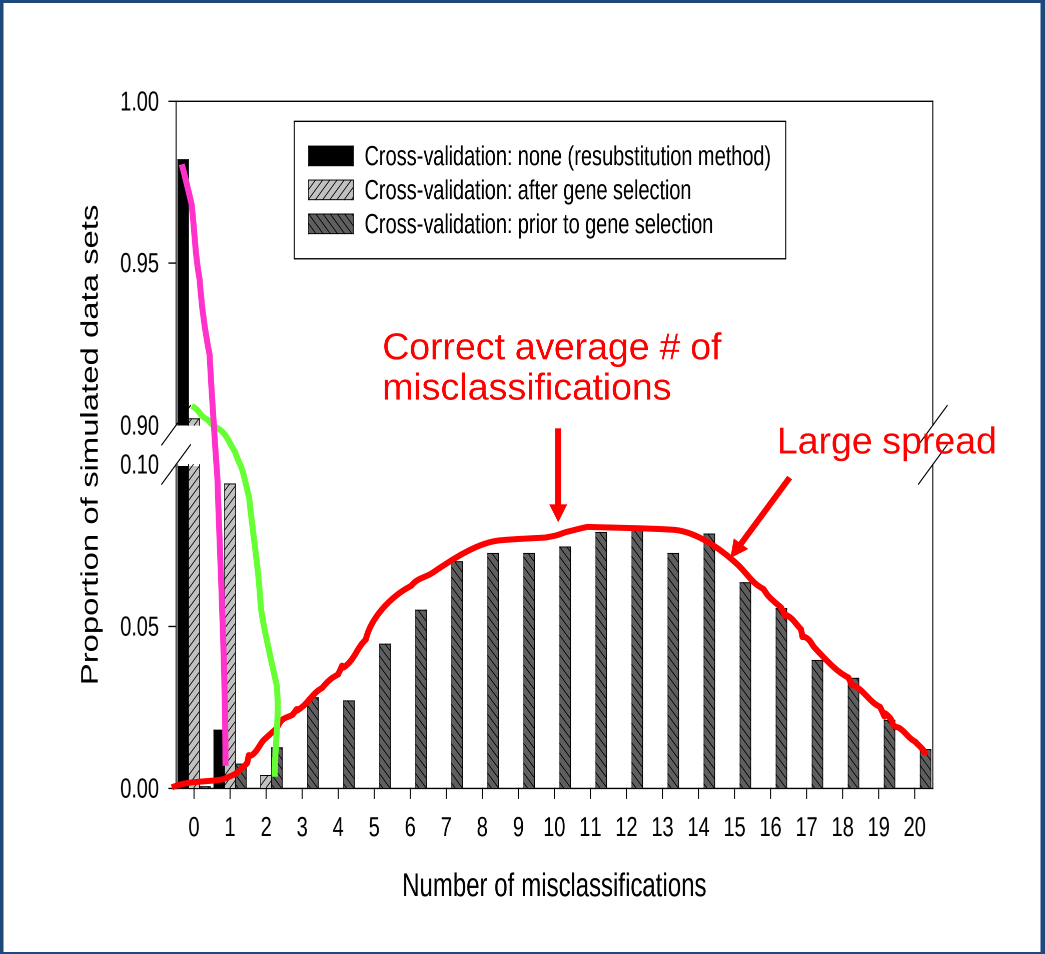
Domain 3: Avoid partial resubstitution

Simulation experiment: 20 specimens; expression levels of 6000 genes randomly generated (Gaussian noise); arbitrary split of specimens into two groups of 10

Prediction Method:

- Compound covariate
- Use 10 most differentially expressed genes to build classifier
- Calculate number of misclassifications

Repeat simulation 2,000 times



Simon et al, *J Natl Cancer Inst* 2003;95:14-18

Domain 3: Avoid combining training & test sets

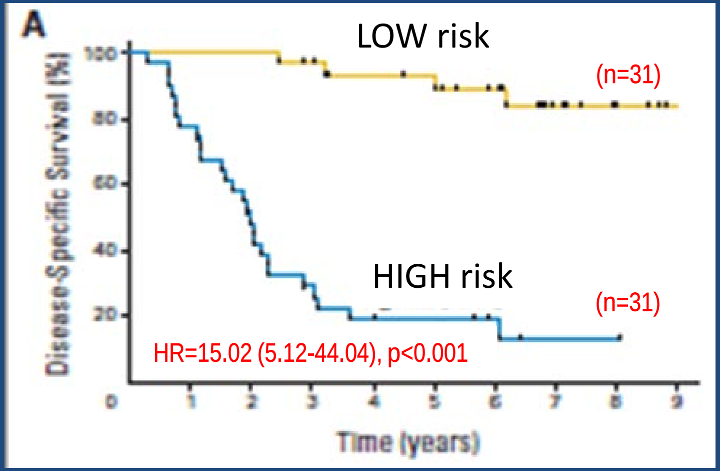
Multivariable Model for Overall Survival
(Training and Test sets combined)

Variable	HR	95% CI	P
Genomic score	2.43	1.94 – 3.06	< 0.001
Stand. molec. factor 1	1.77	1.41 – 2.22	< 0.001
Stand. molec. factor 2	0.66	0.48 – 0.93	0.02
Age group, ≥ 60 yrs vs < 60 yrs	2.22	1.76 – 2.79	< 0.001

Combining Training data (used to develop genomic score) with Test data destroys the validation and interpretability of the adjusted effects

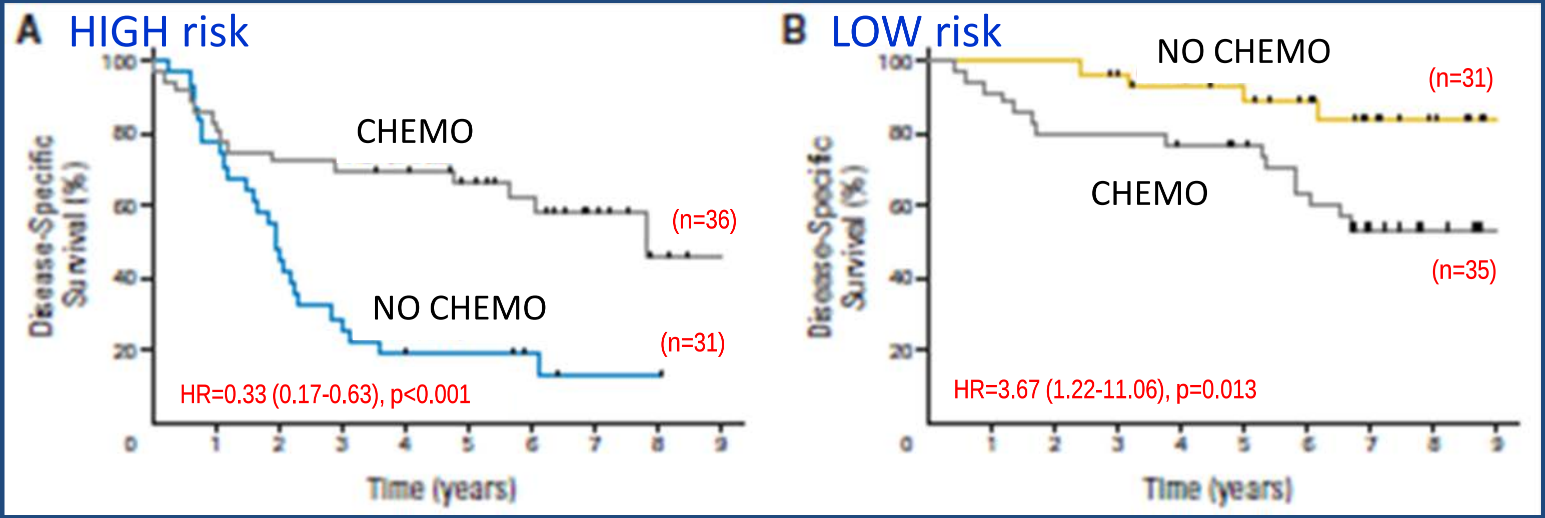
Nowhere in the paper was a multivariate analysis based solely on the Test set presented.

Domain 3: Avoid comparisons with resubstitution estimates



Simon & Freidlin, [Correspondence]
J Natl Cancer Inst 2012;103(5):445

Prognostic classifier fit using gene expression microarray data from clinical trial arm on which patients received no adjuvant chemotherapy (resubstitution)



Does the genomic predictor identify groups of patients who benefit differently from adjuvant chemotherapy?
Can’t conclude anything.

Domain 3: Requirements for a rigorous validation of a predictor

- ❑ The predictor to be tested must be completely **LOCKED DOWN** and there must be a **PRE-SPECIFIED PERFORMANCE METRIC**. The lockdown includes all steps in the data pre-processing and prediction algorithm.
- ❑ The **INDEPENDENT VALIDATION DATA** should be generated from specimens collected at a different time, or in a different place, and according to the pre-specified collection protocol.
- ❑ Assays for the validation specimen set should be run at a different time or in a different laboratory but according to the **IDENTICAL ASSAY** protocol as was used for the training set.
- ❑ The individuals developing the predictor must remain completely **BLINDED** to the validation data.
- ❑ The validation **DATA SHOULD NOT BE CHANGED** based on the performance of the predictor.
- ❑ The **PREDICTOR SHOULD NOT BE ADJUSTED** after its performance has been observed on any part of the validation data. Otherwise, the validation is compromised and a new validation may be required.

Domain 3: Fully-specified "locked down" predictor

Need all of the following:

- ❑ List of individual variables
- ❑ Data pre-processing steps (e.g., normalization/standardization of raw data)
- ❑ Equation/algorithm to make predictions
- ❑ Produces same or highly similar result when *same* data are input multiple times
- ❑ Predictor can be applied *one* case at a time

Domain 3: Examples of predictors *not* locked down

- ❑ Example #1: List of variables (e.g., genes, proteins) with no indication of how to combine the variables
- ❑ Example #2: Data pre-processing using data from a collection of specimens (e.g., each gene expression value is standardized across a collection of cases as $\mathrm{z} = (x - \bar{x})/s$)

  How to pre-process data from a single new case?

  Need to lock down pre-processing parameters or use reference set.

Domain 3: Examples of predictors *not* locked down (cont.)

- Example #3: Use of ranks or percentiles
  - Linear combination scores computed on training set and classified using median score for the training set as cutpoint for classification of the training set cases
  - Linear combination scores computed on test set and classified using median score for the *test* set as cutpoint for classification of the *test* set

Cutpoint may shift from data set to set due to assay batch or cohort effects.

How is a single new case classified?

Domain 3: Example of predictors *not* locked down (cont.)

- Example #4: "Black-box" computer programs that produce varying predictions when run multiple times on same data
  - Stochastic model averaging methods
  - Methods that employ clustering methods with random initial centroids (e.g., some implementations of K-means clustering)

  <u>Example</u>: Same data from ≈100 cases input twice, 20% chance of flipping (low/high risk) prediction, run to run

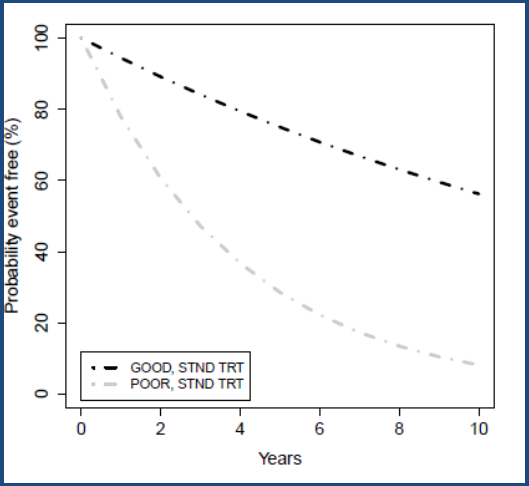
Either varying aspects must be locked (e.g., fix random number seed), or it must be established that variation across repeat runs is minimal.

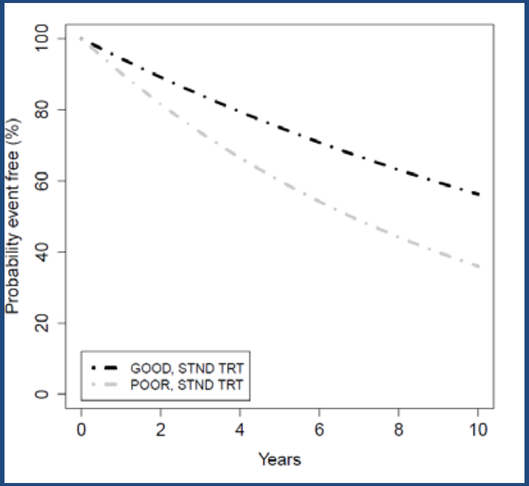
Domain 4: Clinical trial design

- ❑ Clear intended use with clinical utility
- ❑ Is a prospective trial needed, and if so, what design?
- ❑ Protocol with clear objectives, design, statistical analysis plan, locked down predictor
- ❑ Secure database
- ❑ Responsible individuals named

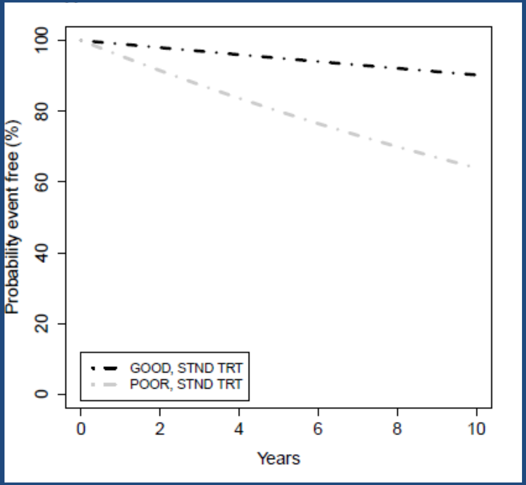
Domain 4: Prognostic ability only sometimes translates to clinical utility

Is this prognostic information helpful ?





Good prognosis group may forego additional therapy



Domain 4: Usually we are more interested in "predictive" ability of a biomarker or omics predictor

- Predictive: Associated with benefit or lack of benefit (potentially even harm) from a particular therapy relative to other available therapy
  - Alternate terms: treatment-selection, treatment-guiding, treatment effect modifier
- Need randomized treatment trial, or at least specimens collected from such a trial

Polley et al, J Natl Cancer Inst 2013;105:1677-1683
McShane & Polley, Clinical Trials 2013; 10: 653-665

Domain 4: Study designs to examine tests/biomarkers for guiding therapy

- ❑ Main types of prospective designs
  - ▪ Biomarker-Enrichment; Biomarker-Strategy; Biomarker-Stratified; All-comers

Sargent D et al. J Clin Oncol 2005;23:2020-2027

Freidlin B et al., J Natl Cancer Inst 2010;102:152-160

Clark G & McShane L, Stat Biopharm Res 2011;3:549-560

- ❑ Prospective-retrospective design
  - ▪ Stored specimens from completed prospective trial
  - ▪ Clear pre-specified study objectives
  - ▪ Rigorous statistical design & analysis plans

Simon et al, *J Natl Cancer Inst* 2009;101:1446-1452

Domain 5: Ethical, legal, and regulatory issues

- ❑ Informed consent discloses investigational use, risks, potential COIs
- ❑ Intellectual property
- ❑ Requirements for tests to be performed in CLIA-certified laboratory
- ❑ Determine if investigational device exemption (IDE) is required from FDA

Case study: Serum proteomic test to guide use of EGFR-TKI therapy for patients with lung cancer

- ❑ Patients with advanced non-small cell lung cancer typically have poor outcome with standard chemotherapies
- ❑ Some new drugs have been designed to be effective against tumors that have alterations in the EGFR gene (EGFR-TKIs)
- ❑ Determination of whether a tumor has an EGFR alteration has traditionally required obtaining a biopsy of the tumor
- ❑ A serum proteomic test, if proven reliable, could avoid the need for tumor biopsy to evaluate likelihood of sensitivity to EGFR-TKIs

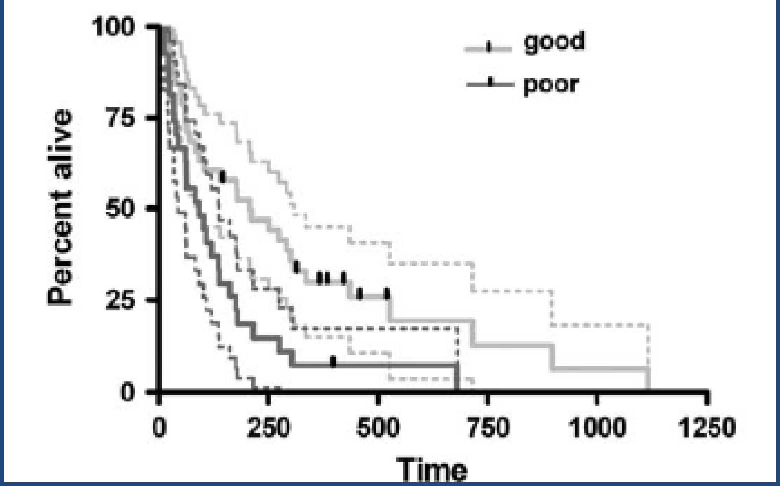
Taguchi et al, *J Natl Cancer Inst* 2007;99:838-46

Model development for serum proteomic test

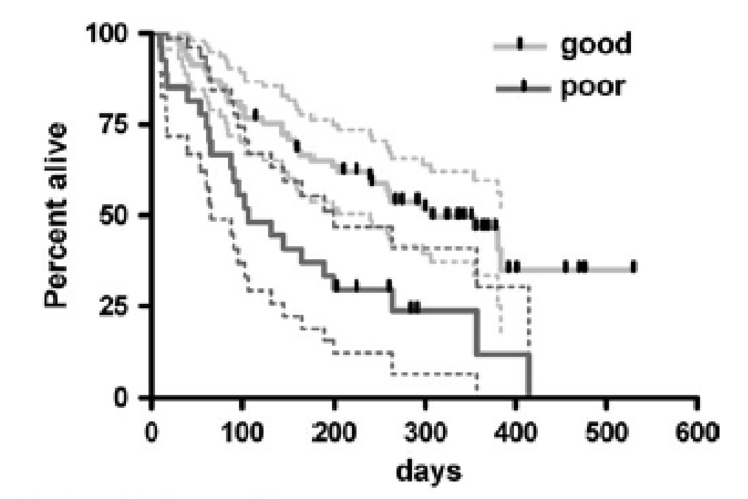
- Serum collected from NSCLC patients before treatment with gefitinib or erlotinib (EGFR-TKIs)
- Analysis by MALDI-MS
- K-nearest neighbor (KNN) algorithm based on 8 distinct m/z features classifies into good or poor outcome
- Training set: n=139 NSCLC patients total from 3 cohorts who received gefitinib
- Preliminary validation cohorts:
  - "Italian B": n=67 sequential patients, late-stage or recurrent NSCLC treated with single-agent gefitinib
  - ECOG 3503: n=96 advanced NSCLC patients treated with first-line erlotinib on single arm Phase II study

Initial assessment of serum proteomic test

Preliminary results for patients treated with EGFR-TKIs



"Italian B": n=67 sequential patients, late-stage or recurrent NSCLC treated with single-agent gefitinib
HR*=0.50, 95% CI=(0.24,0.78), p=0.0054
Median OS
Good: 207 days Poor: 92 days



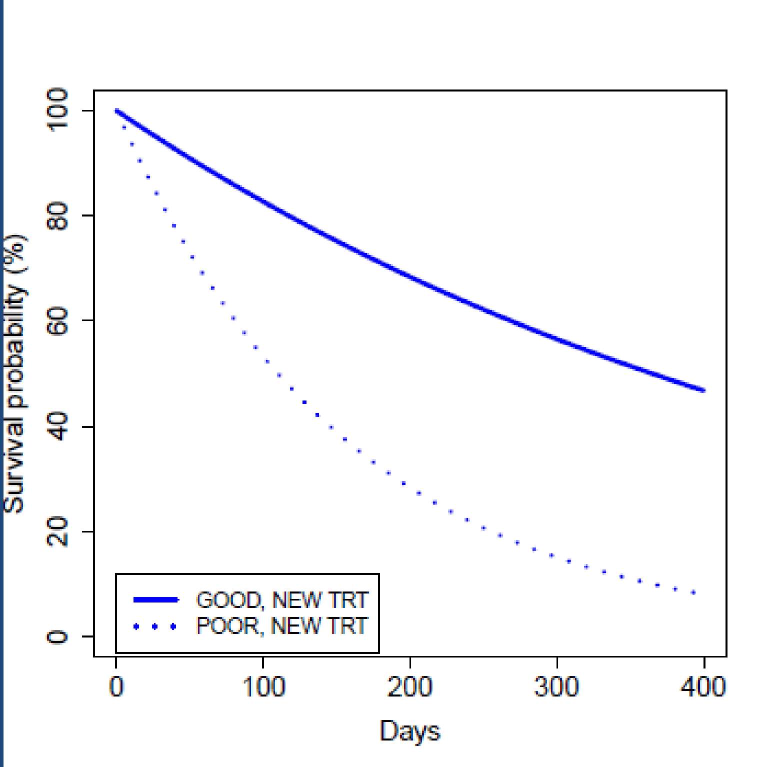
ECOG 3503: n=96 advanced NSCLC patients treated with first-line erlotinib on single arm Phase II study
HR*=0.4, 95% CI=(0.24,0.70), p<0.001
Median OS
Good: 306 days Poor: 107 days

In addition, proteomic test shown to have good analytical reproducibility across 2 labs

*HR for Good:Poor

Serum proteomic test: Predictive or prognostic?

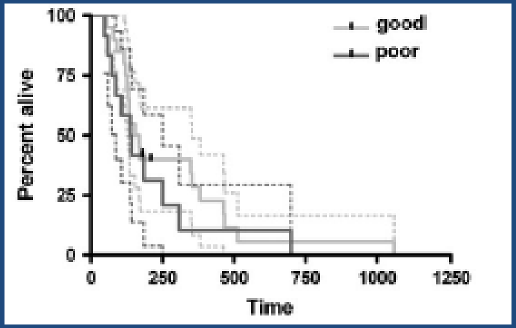
This is what we see for patients who received the EGFR-TKIs



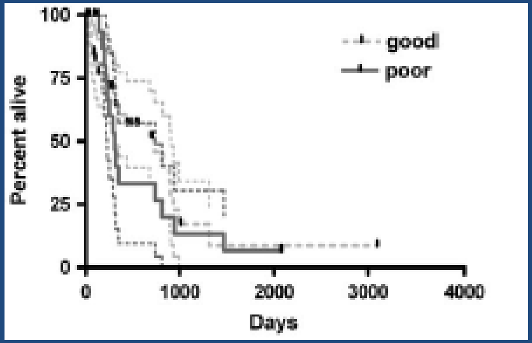
BUT, what does survival look like for patients who receive standard chemotherapy?

Serum proteomic test: Predictive or prognostic?

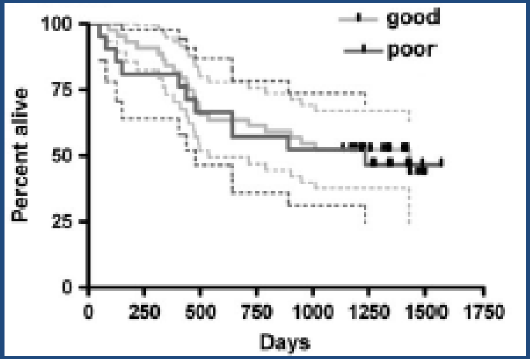
Does test also separate, by outcome, patients who did NOT receive EGFR-TKIs (control cohorts)?



"Italian C": n=32 patients, stage IIIA-IV NSCLC treated with second-line chemotherapy
HR*=0.74, 95% CI=(0.33,1.6), p=0.42
SAME TREND (HR<1) as in EGFR-TKI treated, but not significant



"VU": n=61 patients, advanced NSCLC treated with second-line chemotherapy
HR*=0.81, 95% CI=(0.4,1.6), p=0.54
SAME TREND (HR<1) as in EGFR-TKI treated, but not significant



"Polish": n=65 patients, stage IA-IIB NSCLC treated with second-line chemotherapy
HR*=0.90, 95% CI=(0.43,1.89), p=0.79
SAME TREND (HR<1) as in EGFR-TKI treated, but not significant

*HR for Good:Poor

Serum proteomic test: Need a randomized clinical trial to draw conclusions

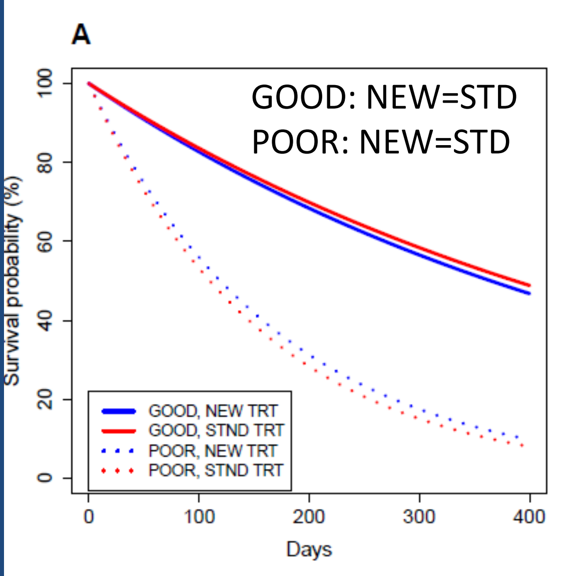
- ❑ Therapy not randomized in the retrospective studies used for development
- ❑ Clinical characteristics (e.g., stage) differed across the patient cohorts

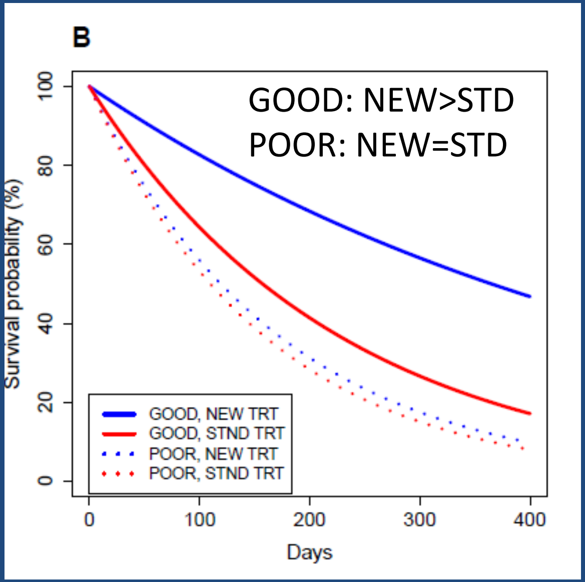
Randomized phase III trial (PROSE) to evaluate ability of serum proteomic test to predict benefit from EGFR-TKIs

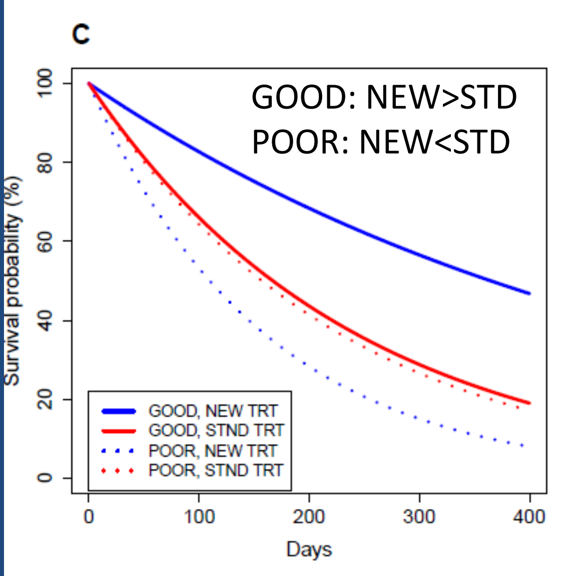
- ❑ Test predictive value of the proteomic test
- ❑ Primary endpoint overall survival (OS)
- ❑ Powered for treatment x proteomic test interaction (biomarker-stratified design)
- ❑ Eligibility
  - ▪ Stage IIIB or IV NSCLC
  - ▪ ≥ 18 years old
  - ▪ Refractory to one prevision platinum-containing regimen
- ❑ Exclusions
  - ▪ Previously received an EGFR-TKI
  - ▪ Uncontrolled brain metastases
  - ▪ Other cardiac, renal, etc. conditions

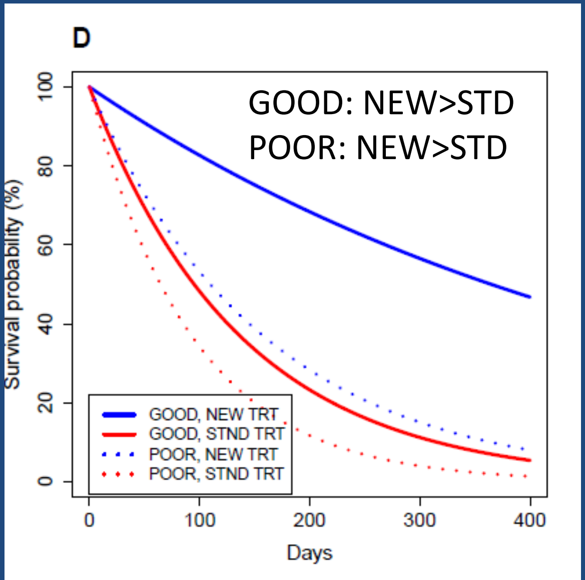
Gregorc et al, *Lancet Oncol* 2014;15:713-721

Serum proteomic test: Some possible clinical trial outcomes

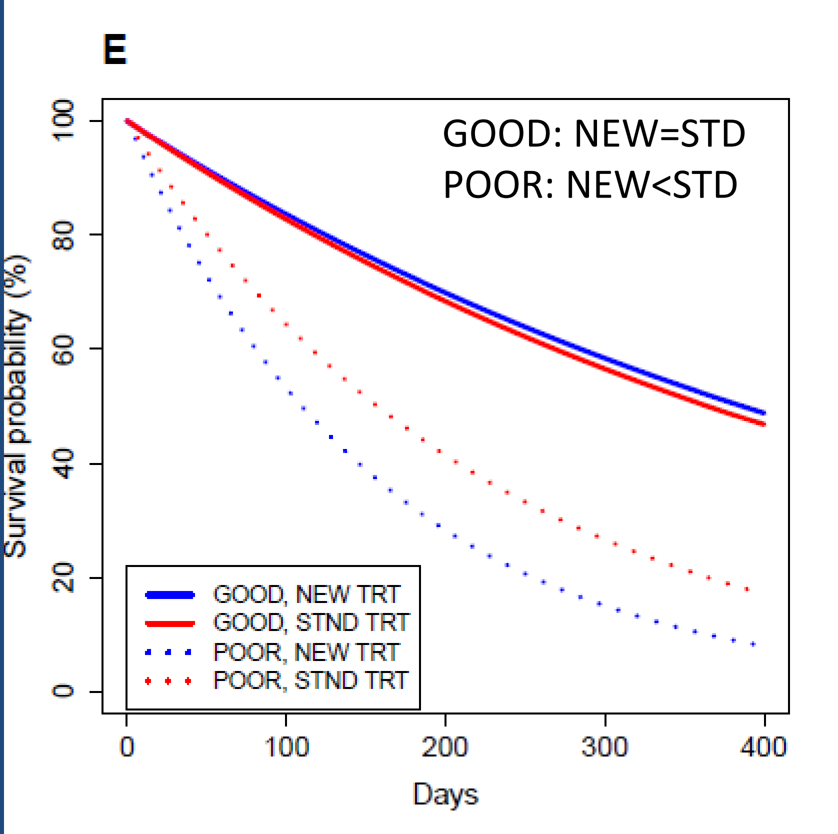




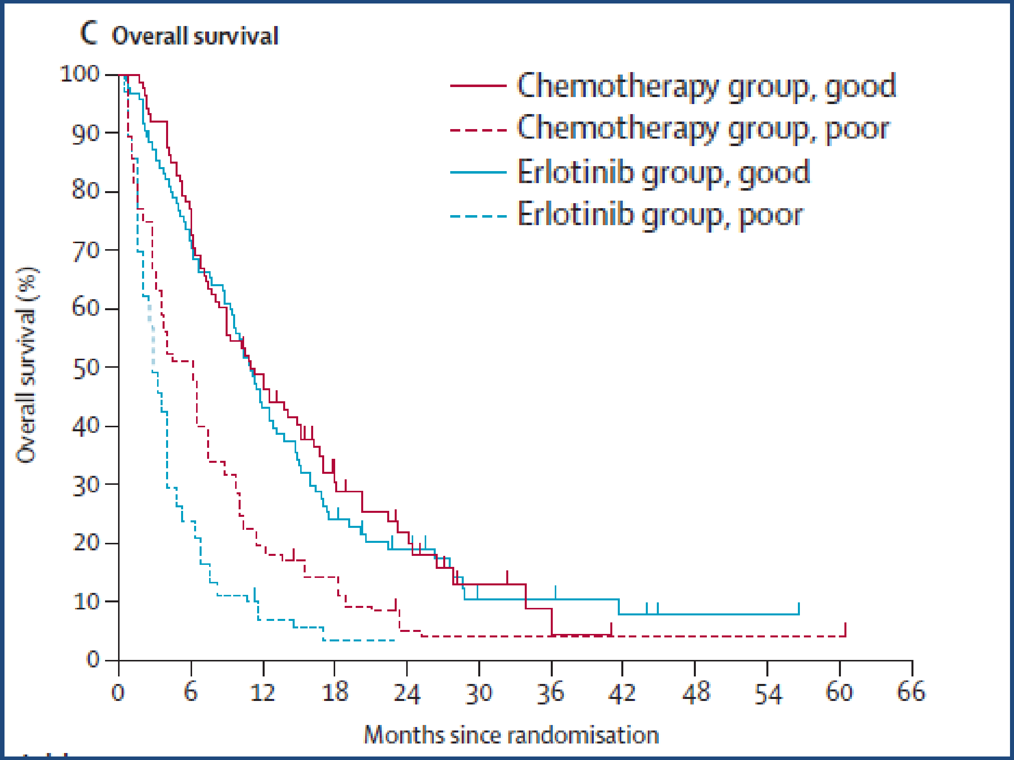




Serum proteomic test: The real clinical trial outcome was none of the above



PROSE trial results for overall survival



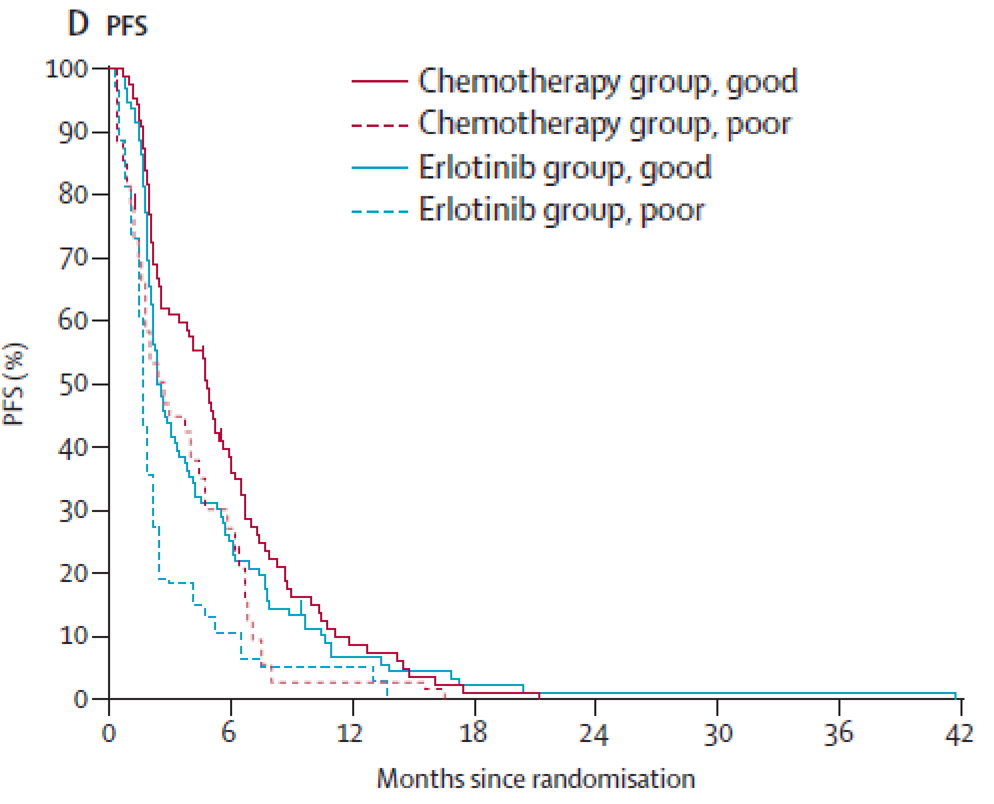
Median Overall Survival (mos.)

	Test result	
Treatment	Good	Poor
Chemo	10.9	6.4
Erlotinib	11.0	3.0
Hazard ratio* (95% CI)	1.06 (0.77-1.46)	1.72 (1.08-2.74)

Interaction p=0.017
*HR for Erlotinib:Chemo

Not even a trend for better outcome with erlotinib in the "good" group.

PROSE trial results for progression-free survival



Median Progression-Free Survival (mos.)

	Test result	
Treatment	Good	Poor
Chemo	4.8	2.8
Erlotinib	2.5	1.7
Hazard ratio* (95% CI)	1.26 (0.94-1.96)	1.51 (0.96-2.38)

Interaction p=0.445

*HR for Erlotinib:Chemo

Trend for longer PFS with chemotherapy in the "good" group.

PROSE trial results

Conclusion drawn by authors:

"Serum protein test status is predictive of differential benefit in overall survival for erlotinib versus chemotherapy in the second-line setting. Patients classified as likely to have a poor outcome have better outcomes on chemotherapy than on erlotinib."

(Gregorc et al, *Lancet Oncol* 2014;15:713-721)

Is this test clinically useful?

Summary remarks

- Scientific teams that develop omics tests should include individuals with statistical expertise
- Familiarize yourself with checklists and reporting guidelines BEFORE you start your study
- Statisticians have a responsibility to engage in the scientific process and not naively churn out statistical analyses

THANK YOU!

lm5h@nih.gov