National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Assessment of omics-based predictor readiness for use in a clinical trial

Lisa Meier McShane
Biometric Research Branch
Division of Cancer Treatment & Diagnosis
U.S. National Cancer Institute

Biopharmaceutical Applied Statistics Symposium
Rockville, MD
November 4, 2014

Disclosures

- I have no financial relationships to disclose.
- I will not discuss off label use and/or investigational use in my presentation.

My perspective

- Statistical/scientific reviewer of NCI-sponsored clinical trials and studies for development and validation of biomarker- and omics-based tests
- Scientific Advisory Board (*Science Translational Medicine*) and Editorial Board (*BMC Medicine*)
- Statistical reviewer for numerous biomedical journals
- Statistical collaborator in research projects involving biomarkers and omics tests

Disclaimers

- The views expressed represent my own and do not necessarily represent views or policies of the National Cancer Institute.
- Examples I cite are all based on true stories or published articles, but I have made minor modifications in some cases to protect identities.

OUTLINE

- Background & definitions
- NCI checklist for readiness of omics-based test to be used in a clinical trial (with emphasis on role of statisticians)
  - Specimens
  - Assays
  - Model development, specification & preliminary performance evaluation
  - Clinical trial design
  - Ethical, legal, and regulatory
- Summary remarks

Working definitions

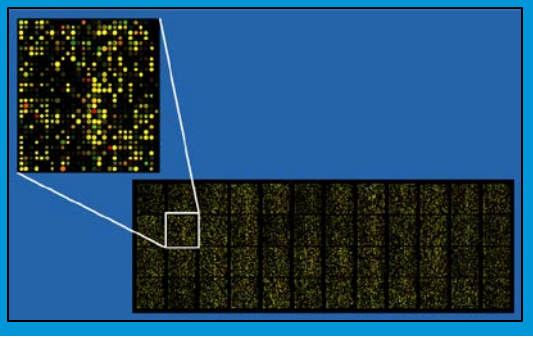
- **Biomarker**
  - http://www.cancer.gov/dictionary: "Biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease."
- **Omics**
  - http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx

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

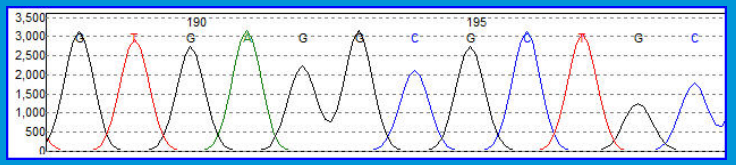
Note: Throughout this talk, biomarkers and omics-based tests will be treated as binary-valued and the two terms will sometimes be used interchangeably for purposes of explaining concepts.

Many examples of omics assays for characterization of biological samples

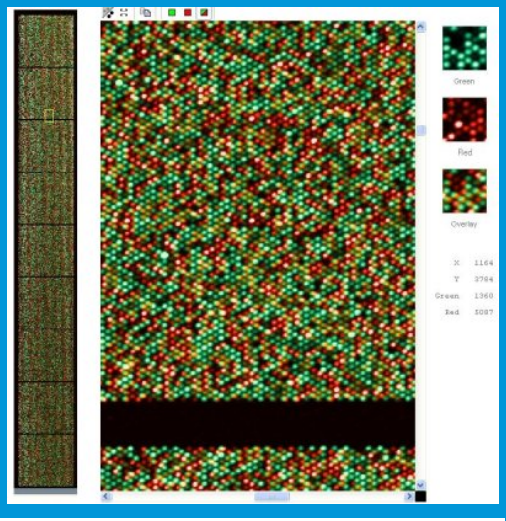
Affymetrix expression GeneChip



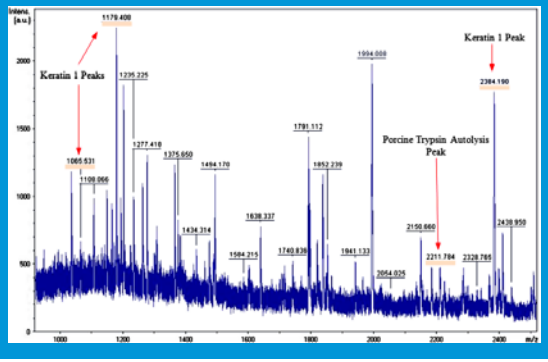
cDNA expression microarray



Mutation sequence surveyor trace



Illumina SNP bead array

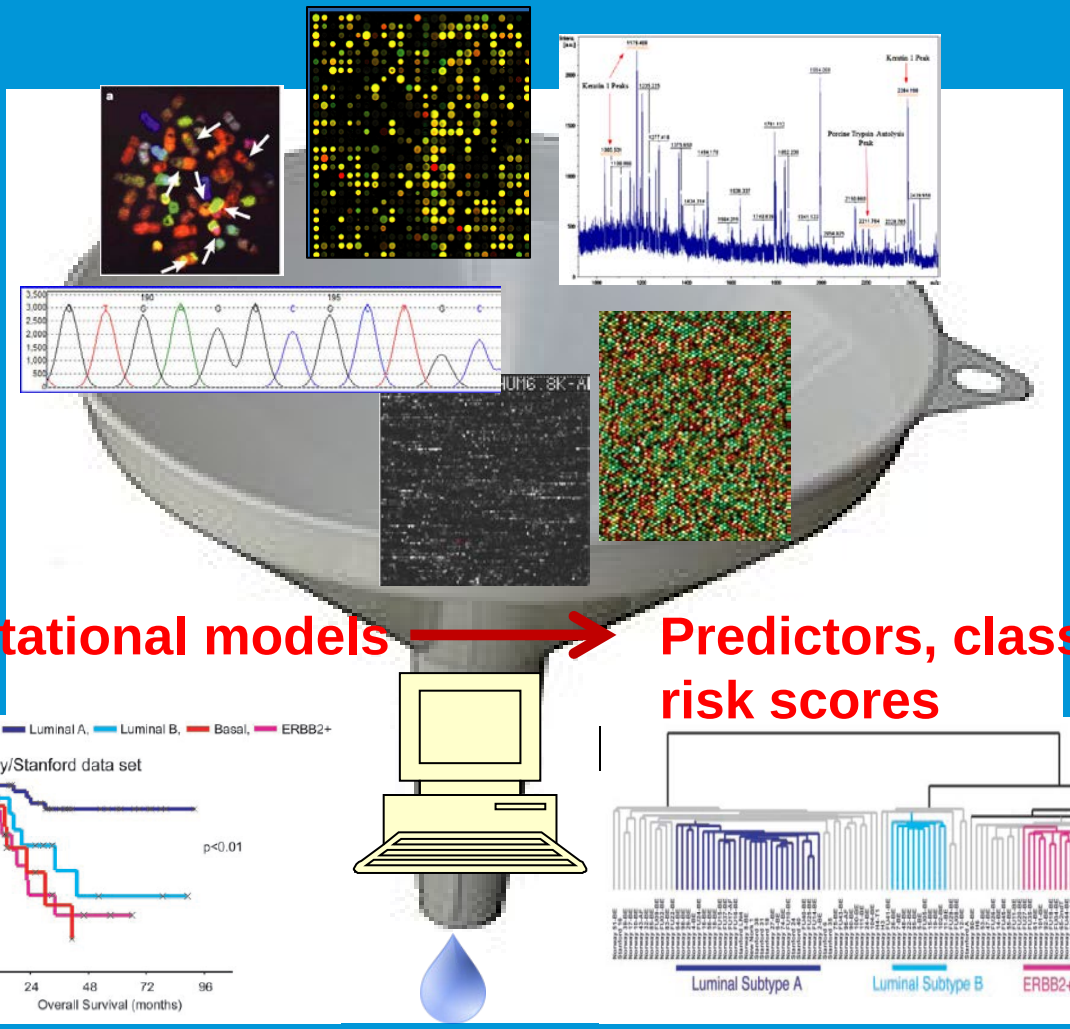


MALDI-TOF proteomic spectrum

Translation from omics discoveries to clinically useful omics-based tests

High-throughput omics assays

Discovery

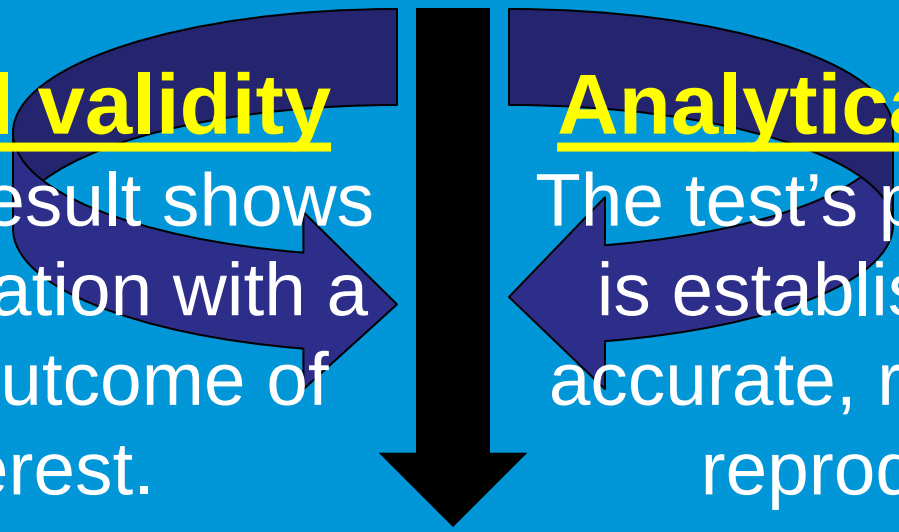


Computational models → **Predictors, classifiers, risk scores**

Clinical Utility?

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

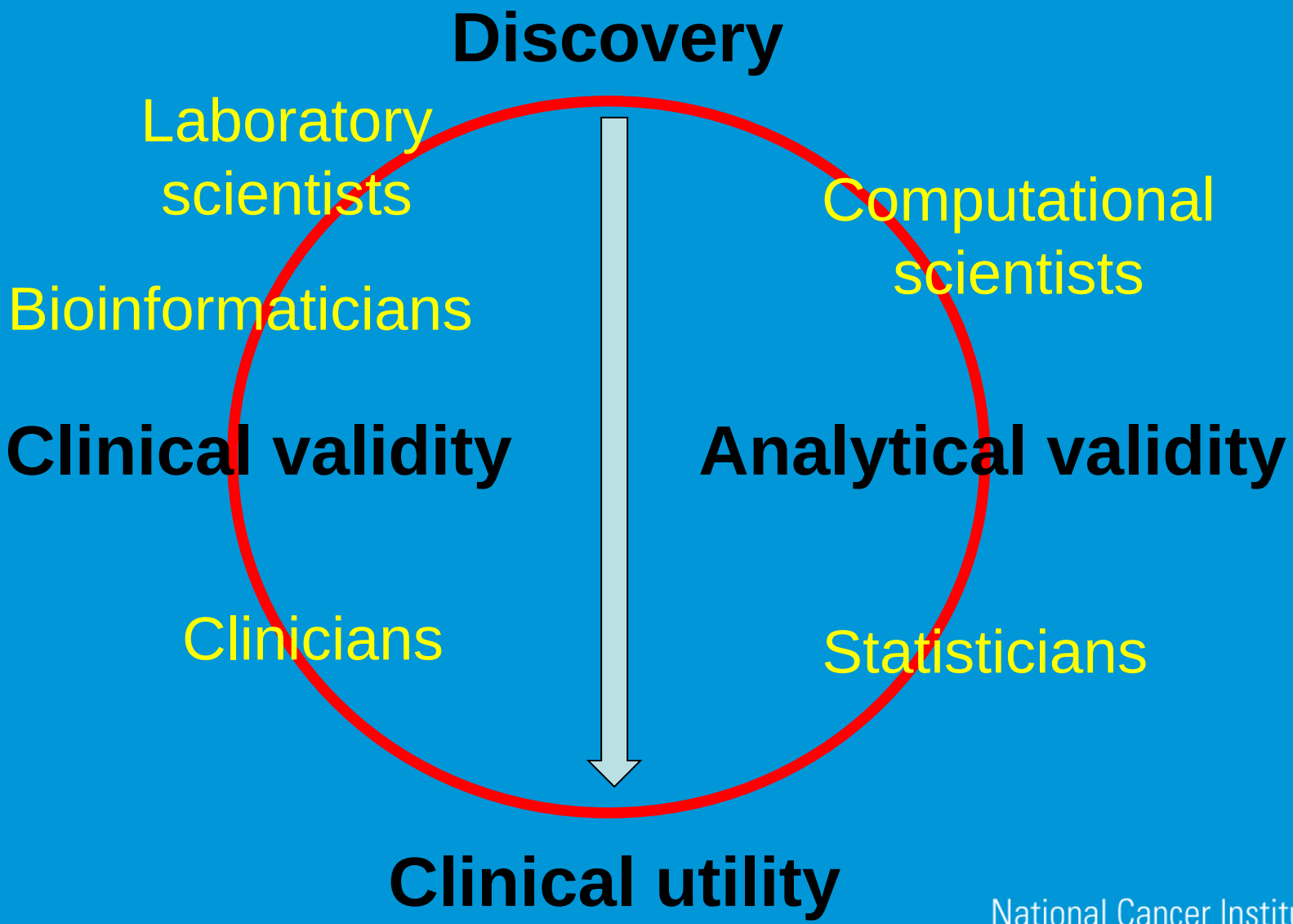
The test's performance is established to be accurate, reliable, and reproducible.

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



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 tests

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

30-point checklist:

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

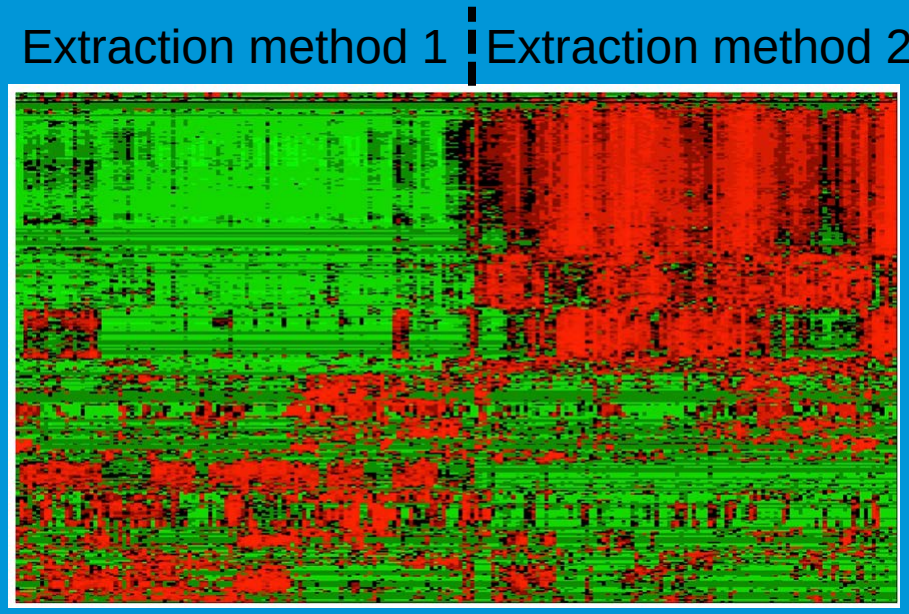
  Jennings et al, *Arch Pathol Lab Med* 2009;133: 743–755
- Quality monitoring
- Turnaround time

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

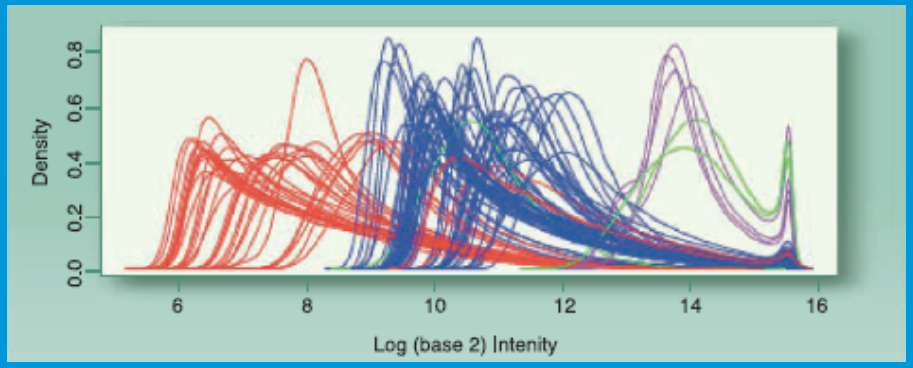
- Quality of data (clinical & omics) used to develop and validate predictor models
- Appropriate statistical approaches for model development and performance assessment
- Intended use - data from clinically relevant patient population

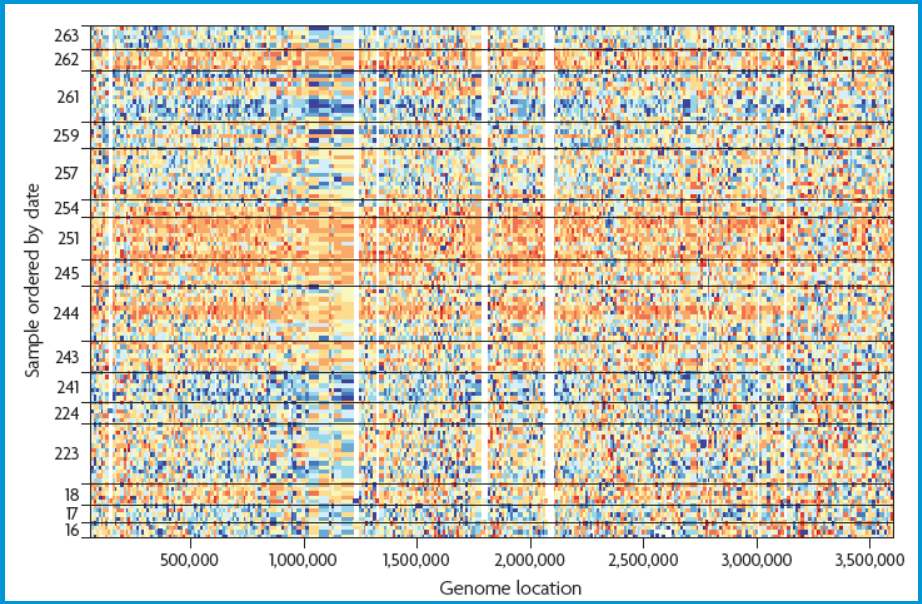
Domain 3: Data quality & batch effects

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

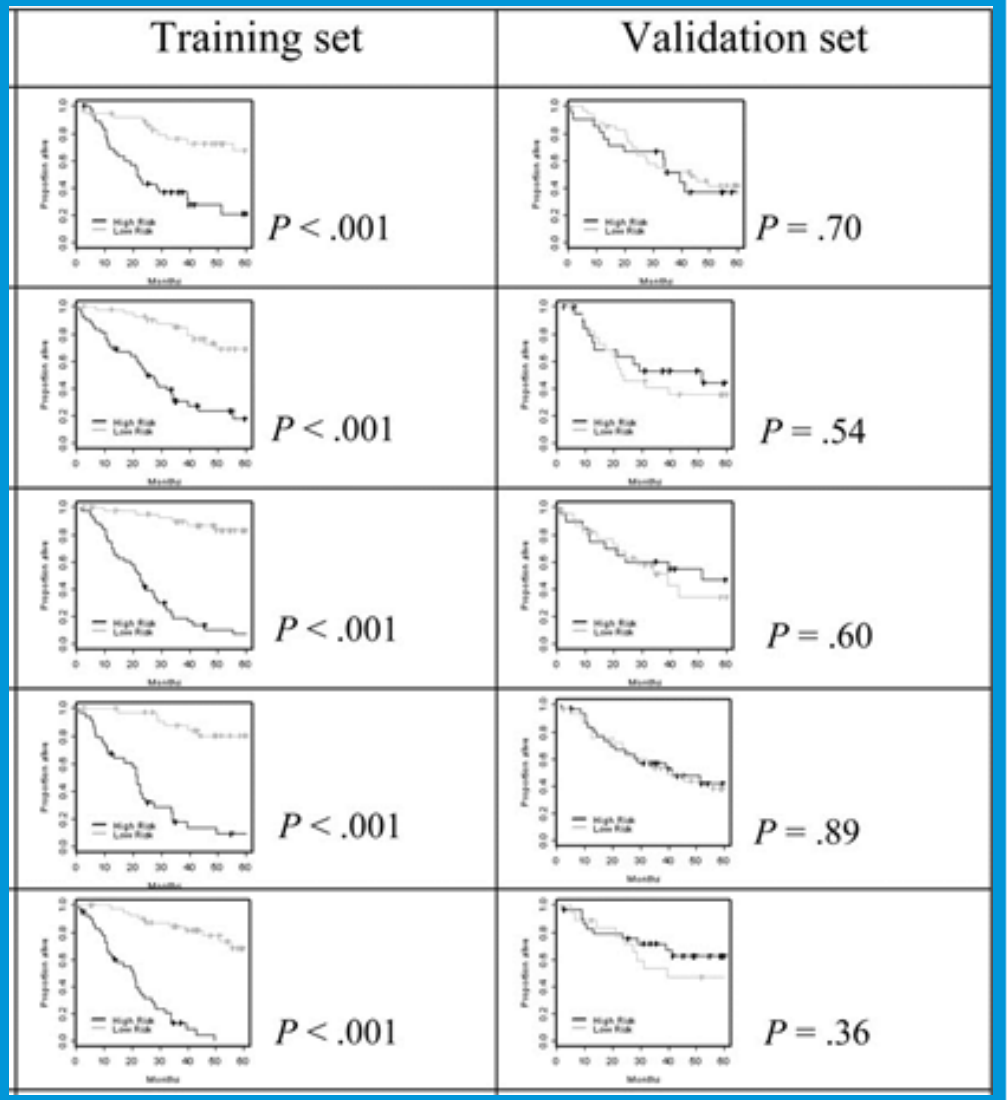
Domain 3: Failure to detect overfitting

- **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

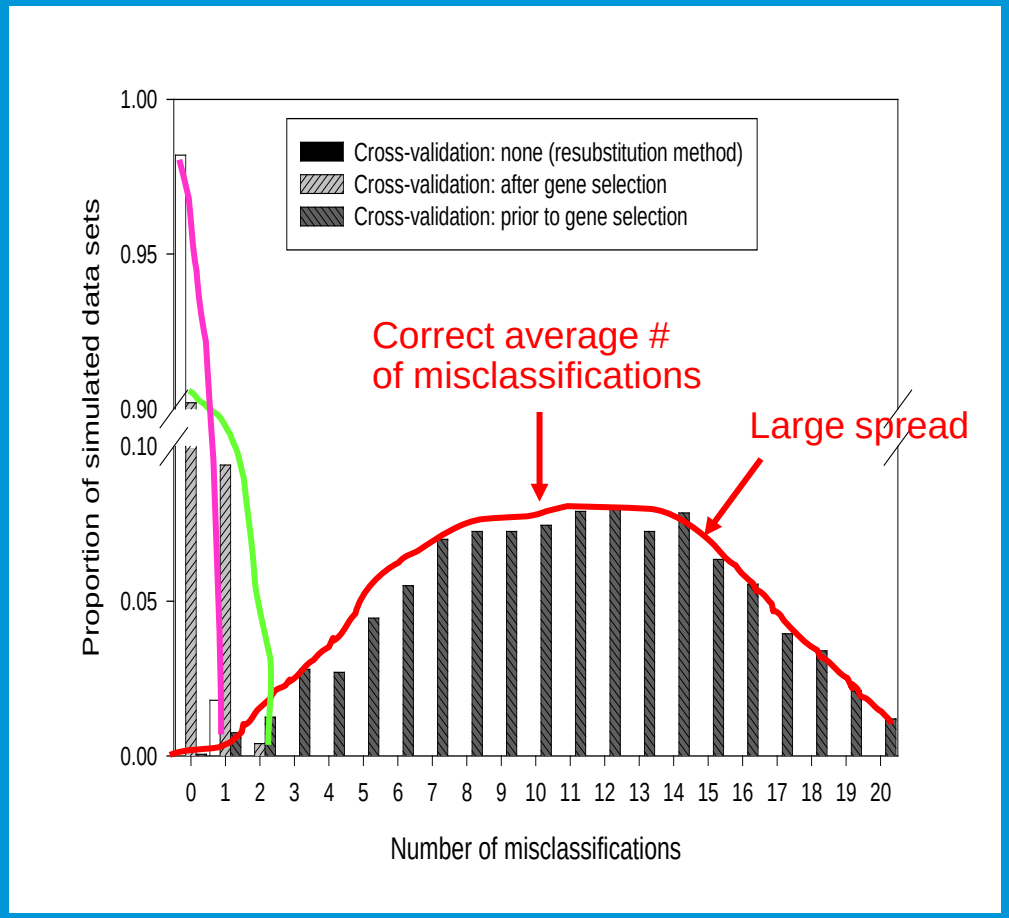
Domain 3: Avoid partial resubstitution

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

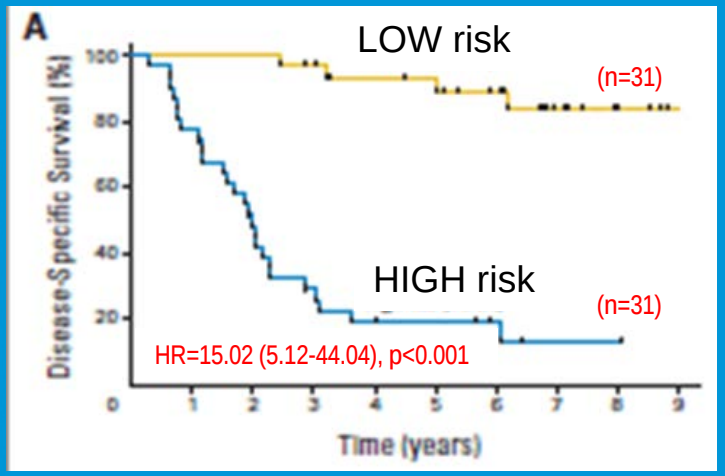
Multivariable Model for Overall Survival
(Training and Test sets combined)

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

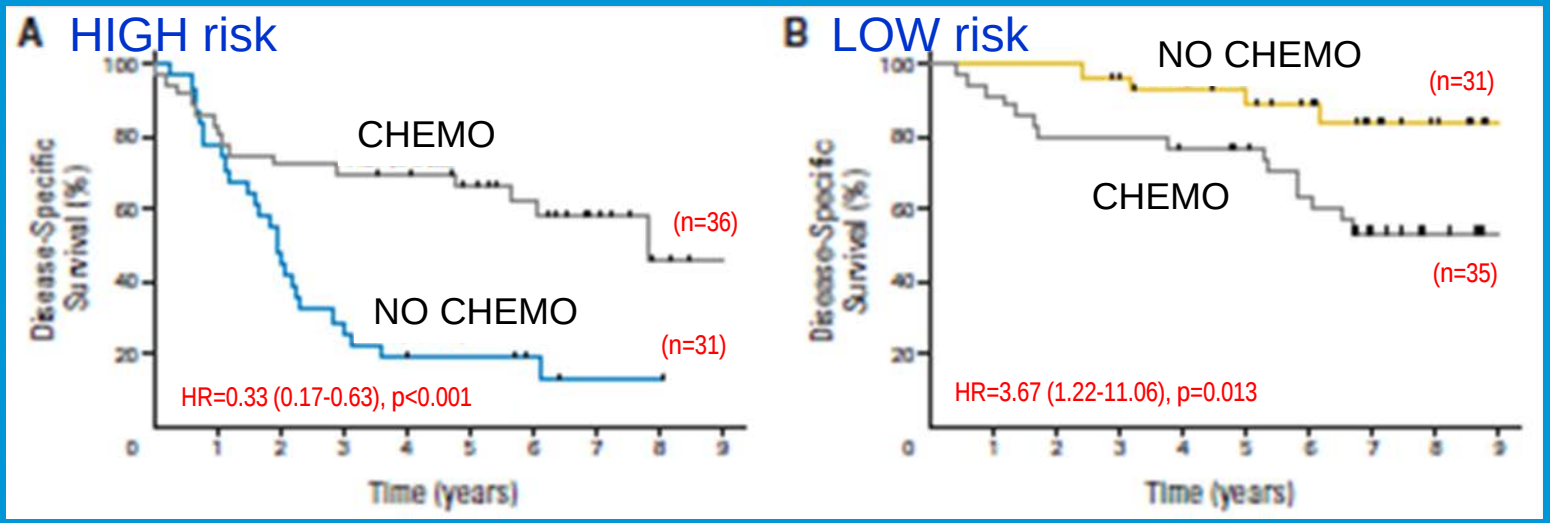
Nowhere in the paper was a multivariate analysis based solely on the Test set presented.

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 $z = (x - \bar{x})/s$)

  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)

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

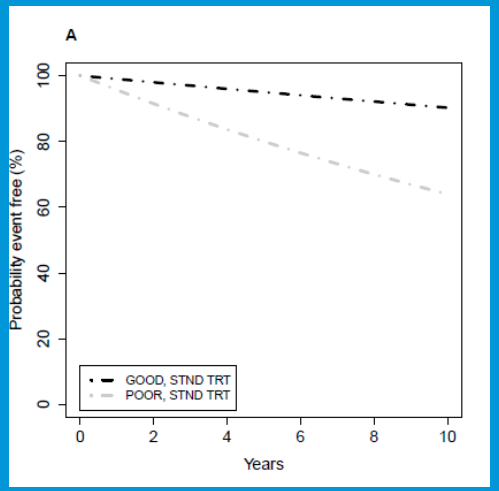
- Clear intended use with clinical utility
- Is a prospective trial needed, and if so, what design?
- Protocol with clear objectives, design, statistical analysis plan, locked down predictor
- Secure database
- Responsible individuals named

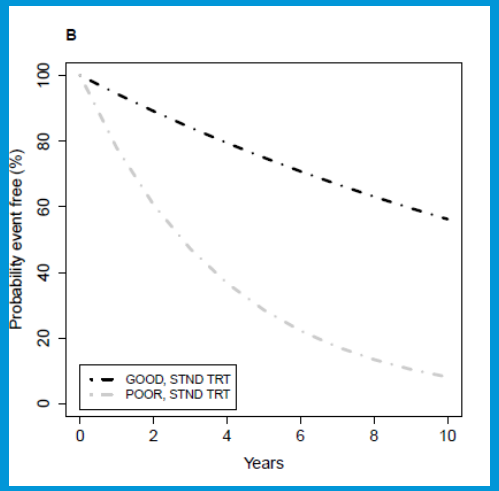
Domain 4: Clinical use – Prognostic

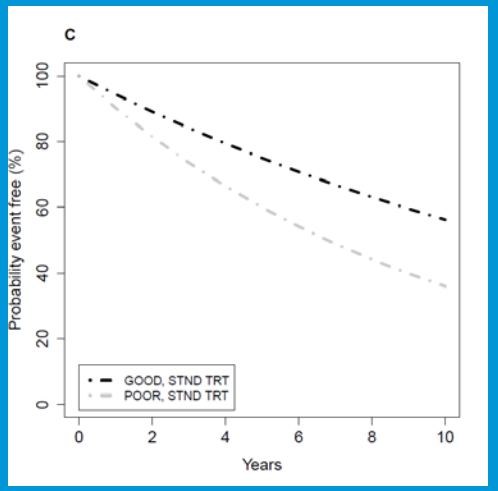
- Associated with clinical outcome in absence of therapy (natural course) *or with standard therapy all patients are likely to receive*
- Not always relevant for therapy decisions

Good prognosis group may forego additional therapy

Is this prognostic information helpful ?







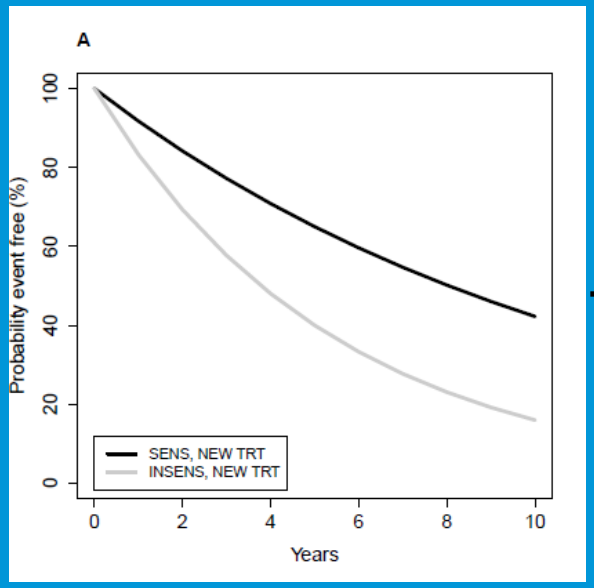
Domain 4: Clinical use – Predictive

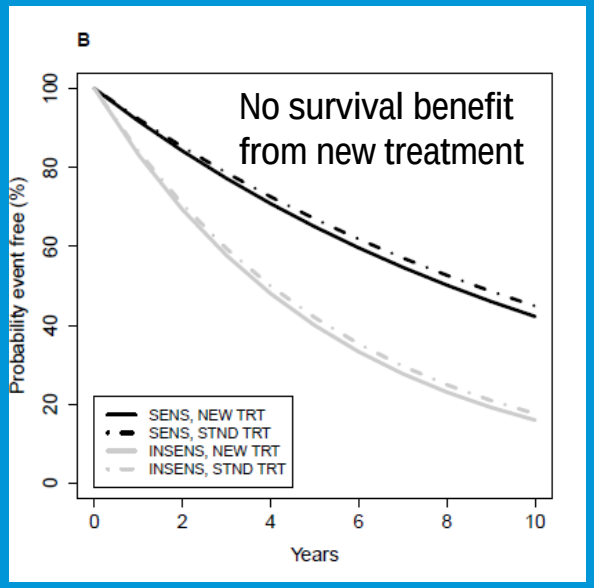
- Associated with benefit or lack of benefit (potentially even harm) from a particular therapy relative to other available therapy
  - Alternate terms: treatment-selection, treatment-guiding, treatment effect modifier
- Generally more useful than prognostic biomarkers for therapeutic decision making

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

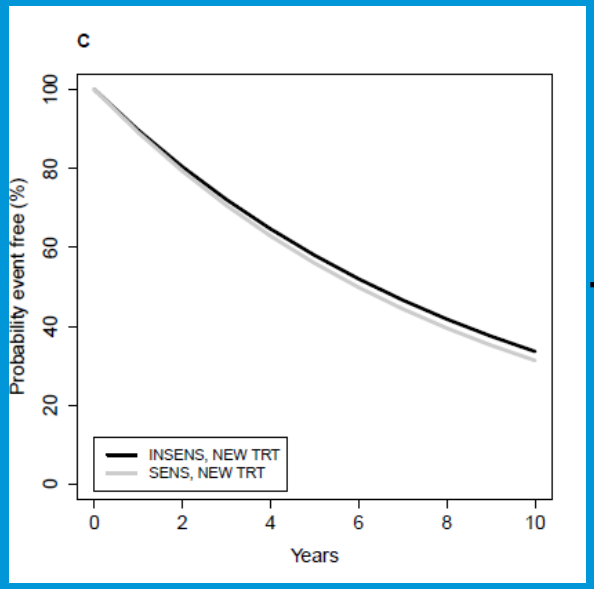
Domain 4: Prognostic vs. predictive

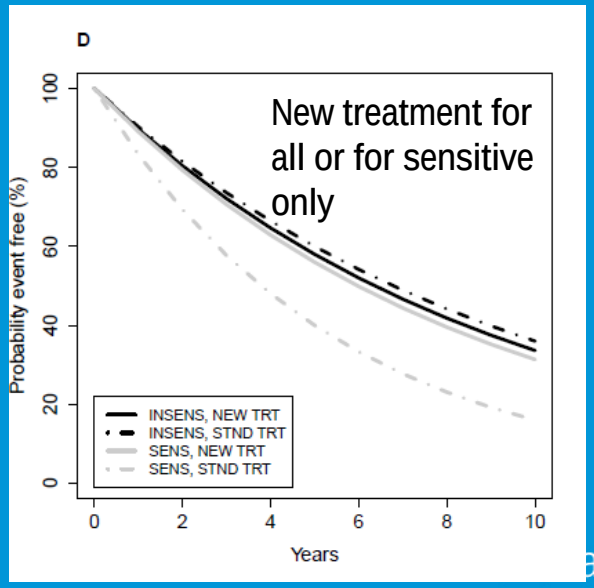
Importance of control groups





Prognostic but not predictive





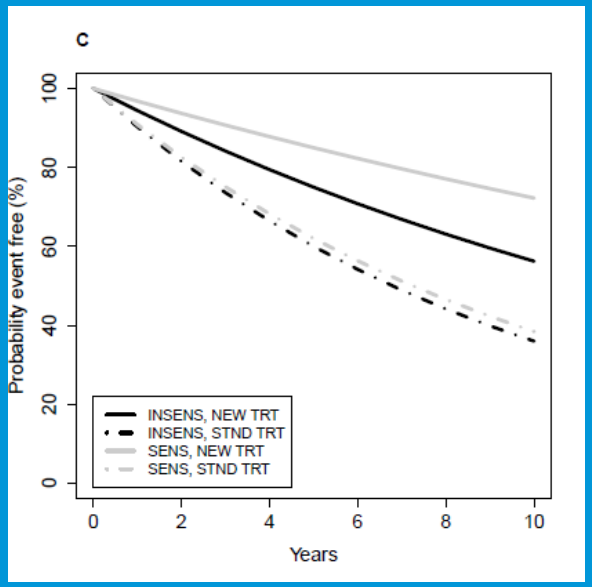
Prognostic and predictive

Domain 4: Statistical language for predictive biomarkers
Treatment-by-biomarker interaction

- Treatment effect (e.g., hazard ratio) varies by biomarker status
  - **QUANTITATIVE** interaction: Treatment benefits all patients but by different amounts
  - **QUALITATIVE** interaction: Patients "positive" for the biomarker benefit from the treatment while others receive no benefit or possibly even harm

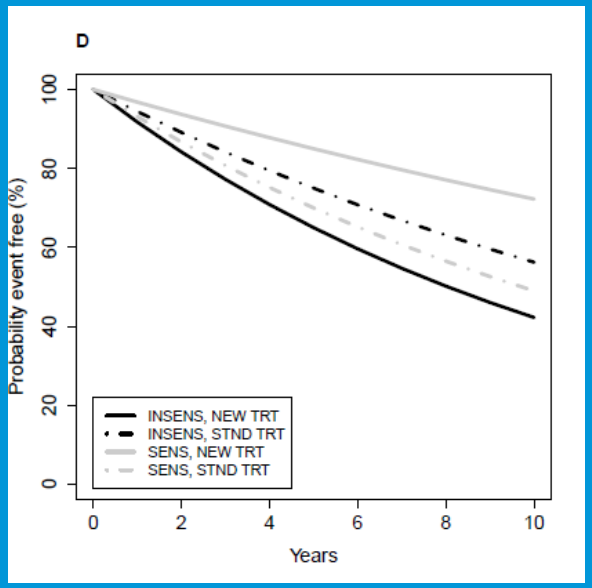
Domain 4: More on predictive tests

Quantitative versus qualitative interaction



Quantitative interaction

Both “sensitive” and “insensitive” subgroups benefit from new therapy, but by different amounts



Qualitative interaction

“Sensitive” subgroup has better outcome on new therapy compared to standard, but “insensitive” subgroup has better outcome on standard therapy.

Domain 4: Clinical trial design

- Main types of prospective designs
  - Biomarker-Enrichment
  - Biomarker-Strategy
  - Biomarker-Stratified

  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
  - Use of stored specimens from a 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 us 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

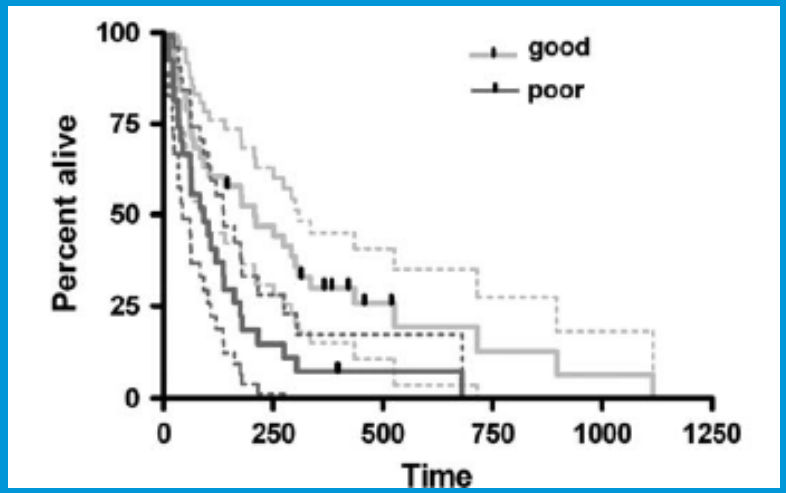
Taguchi et al, *J Natl Cancer Inst* 2007;99:838-46

Model development for serum proteomic test

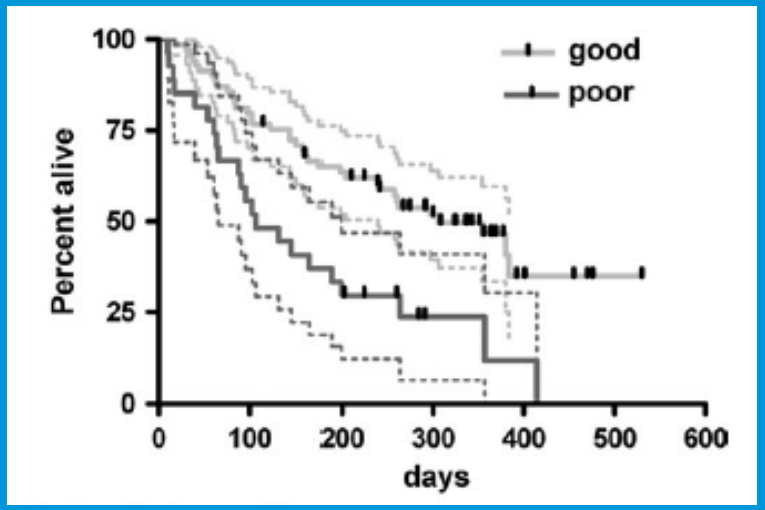
- Serum collected from NSCLC patients before treatment with gefitinib or erlotinib (EGFR-TKIs)
- Analysis by MALDI-MS
- K-nearest neighbor (KNN) algorithm based on 8 distinct m/z features classifies into good or poor outcome
- Training set: n=139 NSCLC patients total from 3 cohorts who received gefitinib
- 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



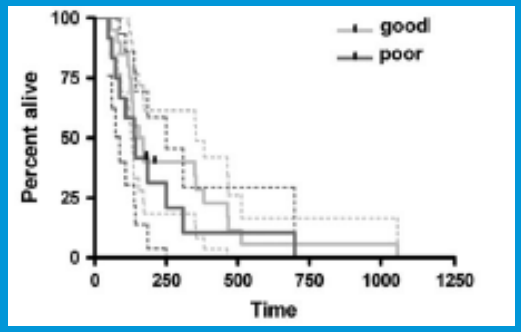
ECOG 3503: n=96 advanced NSCLC patients treated with first-line erlotinib on single arm Phase II study
HR*=0.4, 95% CI=(0.24,0.70), p<0.001
Median OS
Good: 306 days Poor: 107 days

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

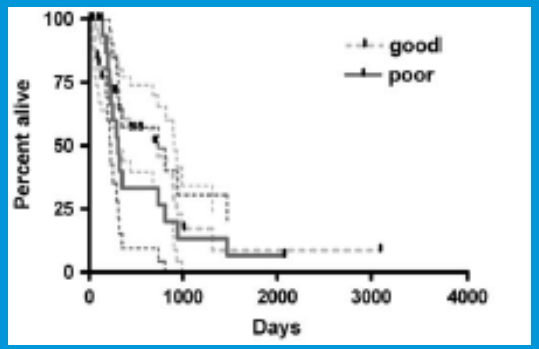
*HR for Good:Poor

Serum proteomic test: Predictive or prognostic?

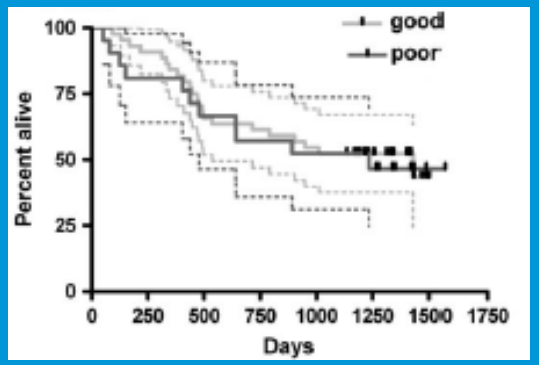
Does test also separate, by outcome, patients who did NOT receive EGFR-TKIs (control cohorts)?



"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

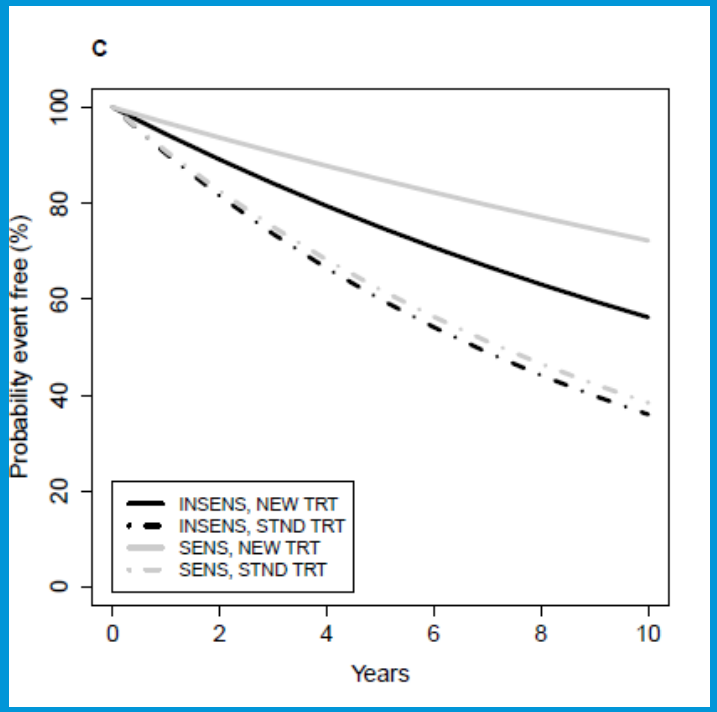


"Polish": n=65 patients, stage IA-IIB NSCLC treated with second-line chemotherapy
HR*=0.90, 95% CI=(0.43,1.89), p=0.79
SAME TREND (HR<1) as in EGFR-TKI treated, but not significant

*HR for Good:Poor

Another look at prognostic vs. predictive

- If prognostic effect of predictor, good:poor (SENS:INSENS), is different between new and standard therapy settings, then there is an interaction but not necessarily *qualitative*.
- Further, if treatment is not randomized it may be difficult to conclude anything.

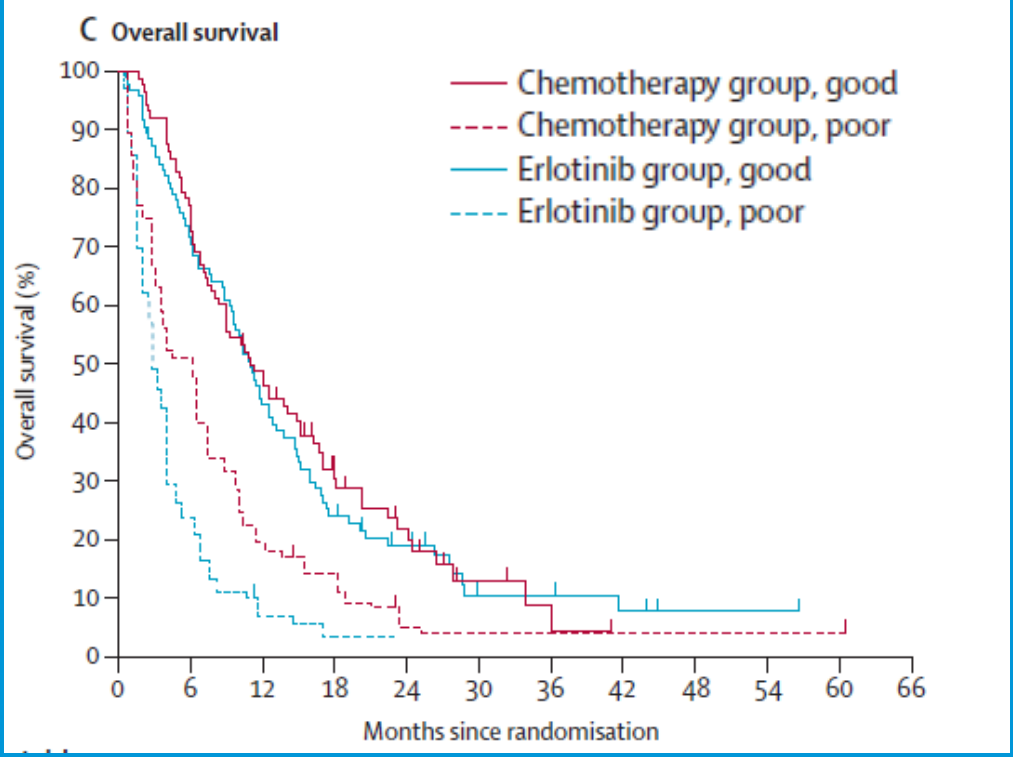


Randomized phase III trial (PROSE) to evaluate ability of serum proteomic test to predict benefit from EGFR-TKIs

- 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

PROSE trial results for overall survival



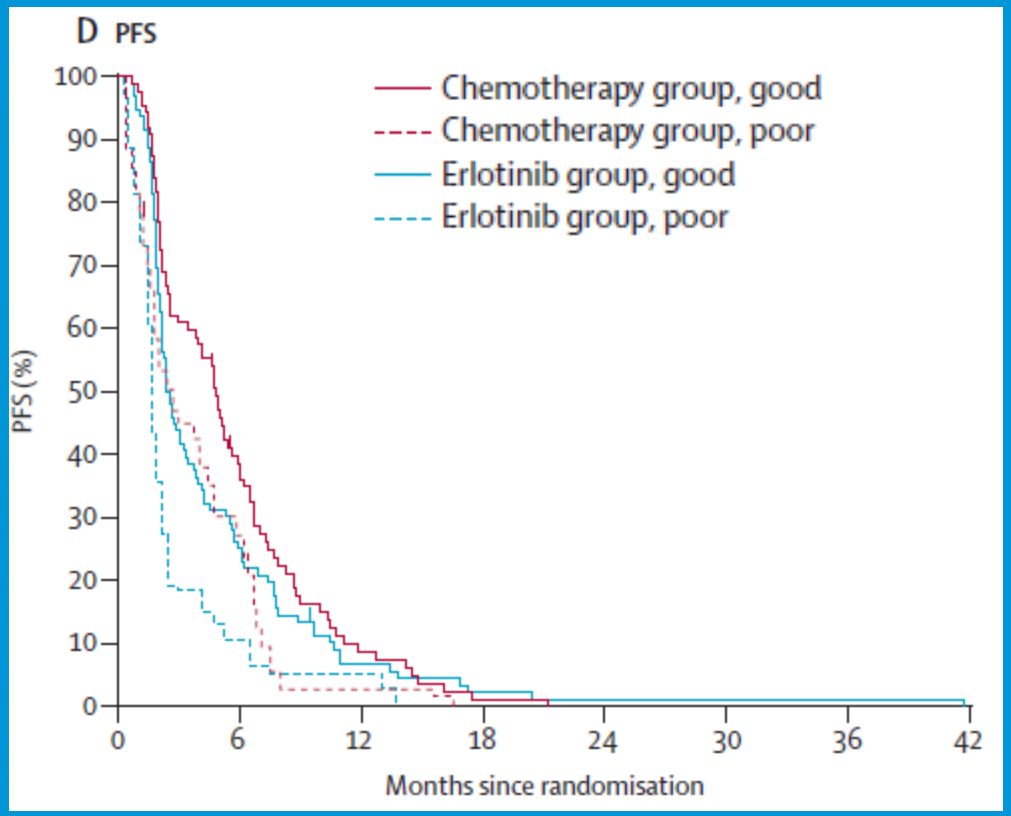
Median Overall Survival (mos.)

	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

Not even a trend for better outcome with erlotinib 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 consistent with the pre-validation?
Was this the pre-specified hypothesis?
How would this be used clinically?

Summary remarks

- Scientific teams that develop omics tests should include individuals with statistical expertise
- Statisticians have responsibility to engage in the scientific process and not naively churn out statistical analyses
- Best practices expected for therapeutics development should be applied to development of omics tests that will guide clinical decisions for patients

THANK YOU!

lm5h@nih.gov