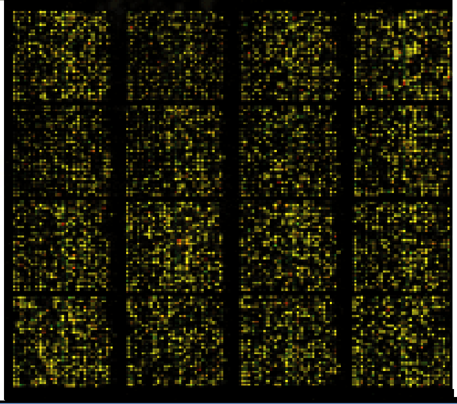
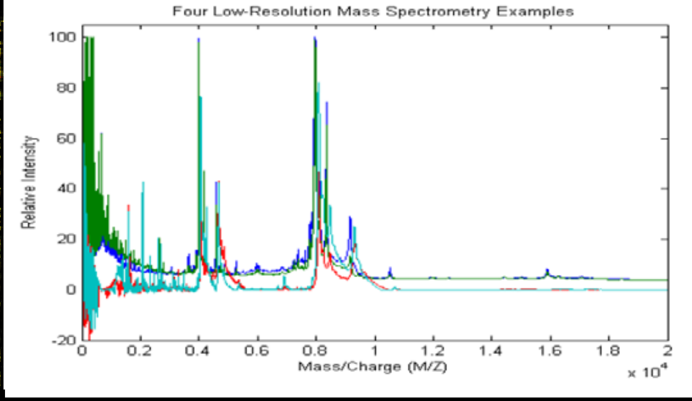
National Cancer Institute





Designing studies to evaluate biomarkers for clinical applications

Presentation to IOM Genomics Workshop:

Evidence for Clinical Utility of Molecular Diagnostics in Oncology

May 24, 2012

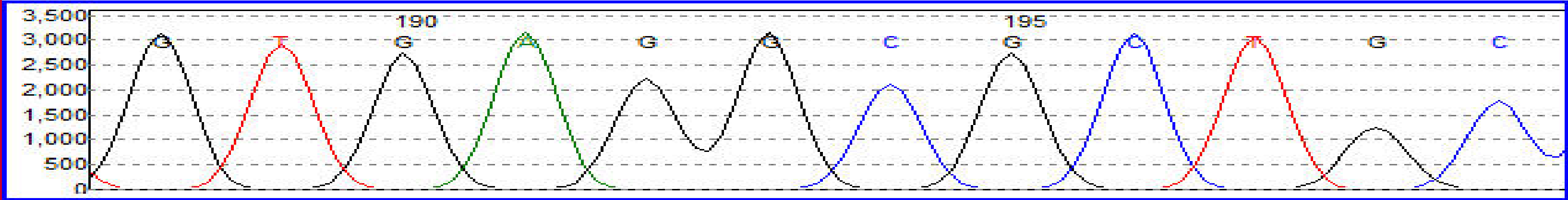
Lisa M. McShane, PhD

Biometric Research Branch

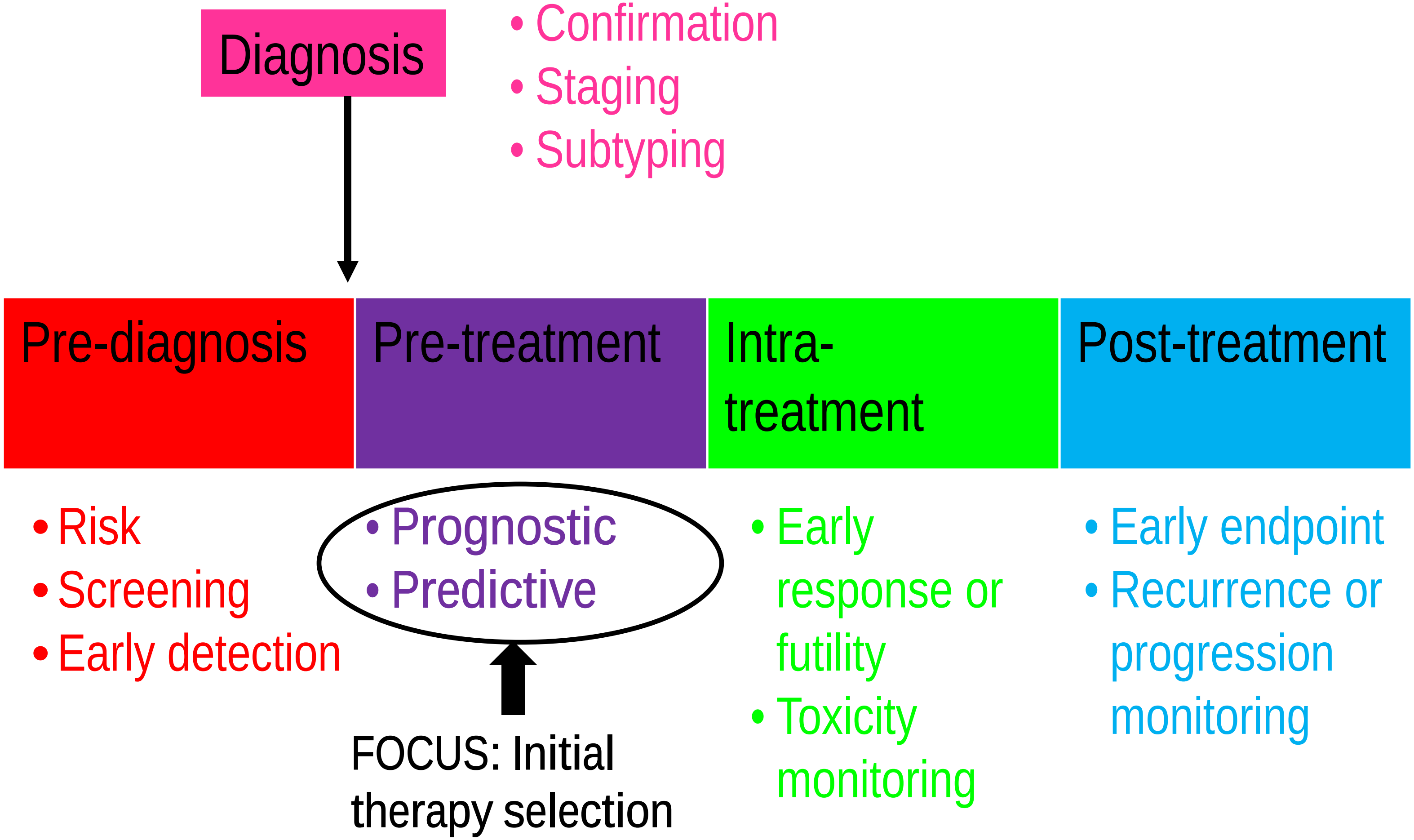
Division of Cancer Treatment and Diagnosis, NCI

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health



Potential roles for molecular diagnostics in medicine



Diagnosis

- Confirmation
- Staging
- Subtyping

Pre-diagnosis

- Risk
- Screening
- Early detection

Pre-treatment

- Prognostic
- Predictive

FOCUS: Initial therapy selection

Intra-treatment

- Early response or futility
- Toxicity monitoring

Post-treatment

- Early endpoint
- Recurrence or progression monitoring

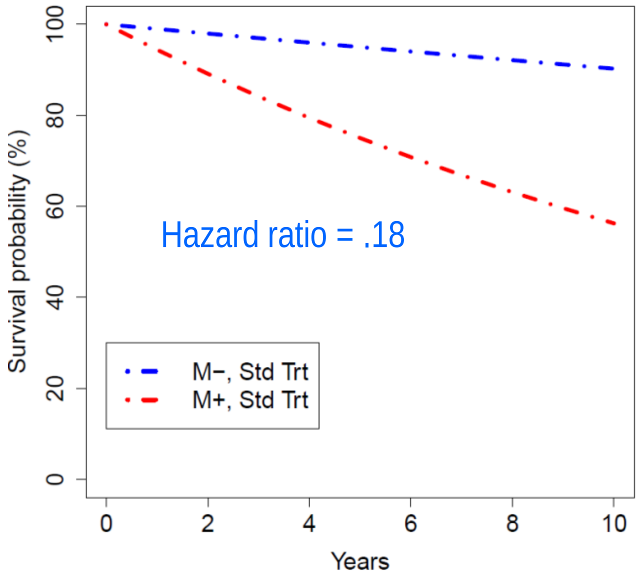
Prognostic & predictive molecular signatures

- Prognostic: Signature associated with clinical outcome in absence of therapy (natural course) *or with standard therapy all patients are likely to receive*
    - Treatment vs. no treatment following surgery
    - Aggressiveness of treatment
    - Examples: OncotypeDX or Mammaprint
- Predictive: Signature associated with benefit or lack of benefit (potentially even harm) from a particular therapy relative to other available therapy
    - Select one treatment vs. another treatment
    - Alternate terms: treatment-selection, treatment-guiding, treatment effect modifier
    - Examples: ER/endocrine therapy, Kras/anti-EGFR mAb

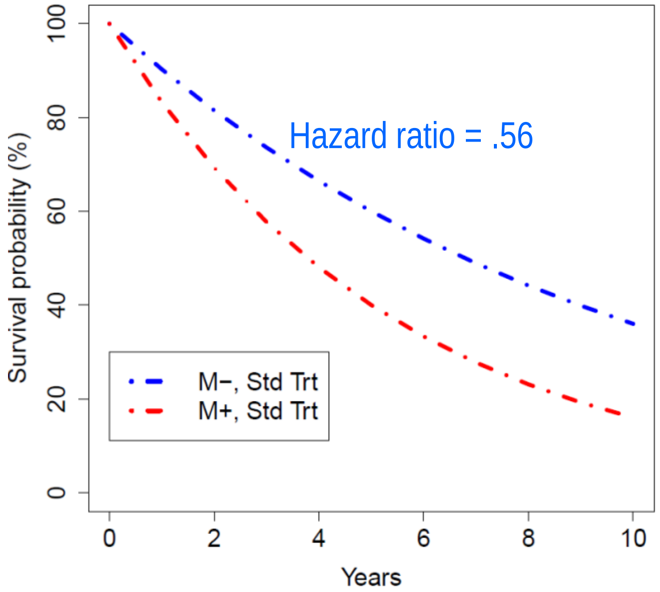
When is a prognostic test clinically useful?

- Is the prognostic information sufficiently strong to influence clinical decisions?
- Does the biomarker provide information beyond standard prognostic factors?
- Does use of the test result in clinical benefit?

Good prognosis group (M-) may forego additional therapy

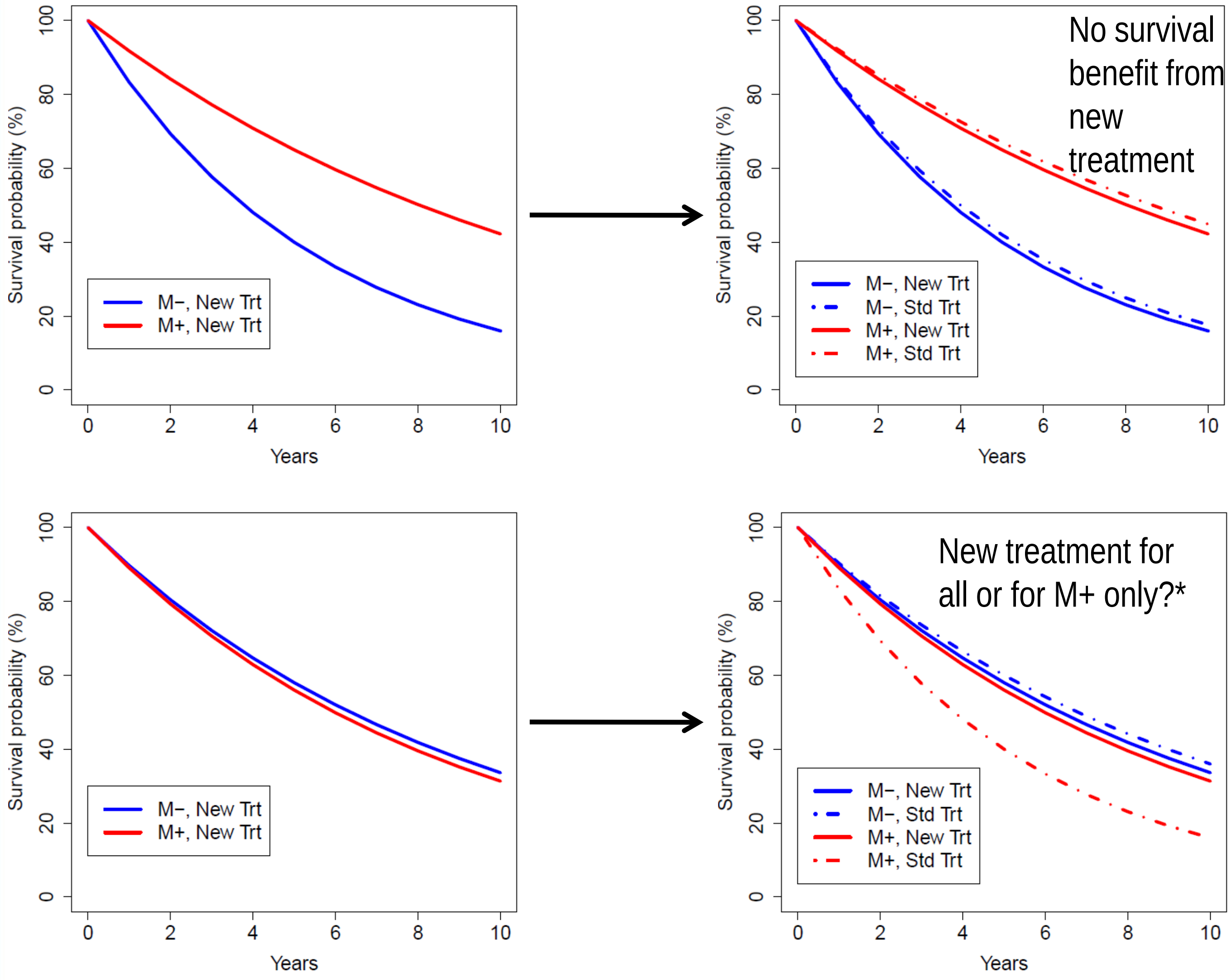


Is this prognostic information helpful?



National Cancer Institute

Prognostic vs. predictive distinction: Importance of control groups



No survival benefit from new treatment

Prognostic but not predictive

New treatment for all or for M+ only?*

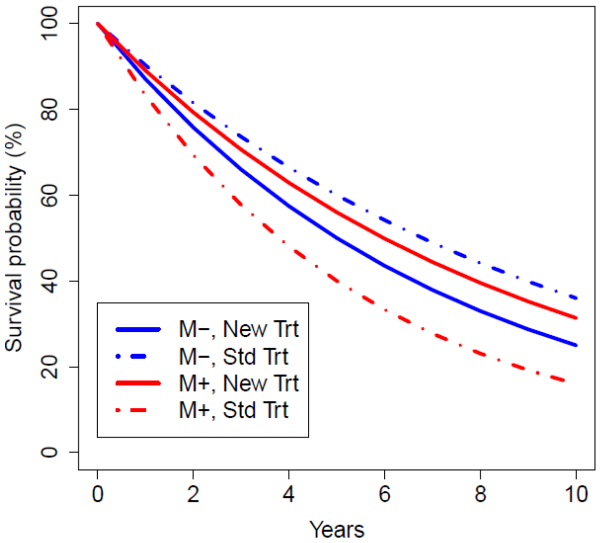
Prognostic and predictive

(*Different considerations might apply for Standard Treatment ± New Treatment)

When is a predictive test clinically useful?

Treatment-by-biomarker interaction: Is it sufficient?

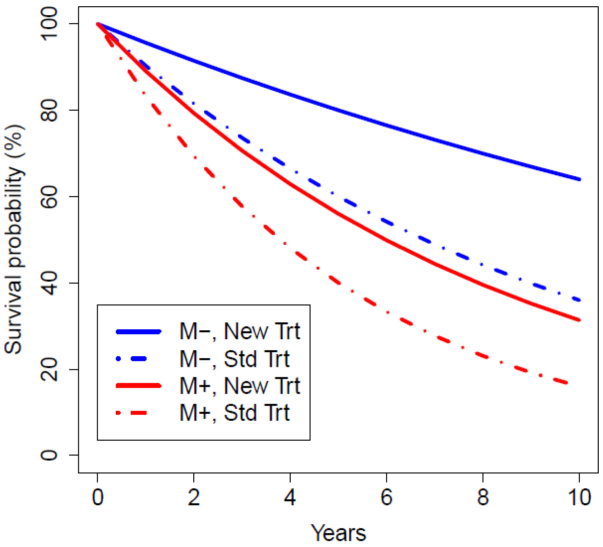
Prognostic and predictive;
New treatment for M+ only



Qualitative interaction

- Std Trt better for M– (HR_= 1.36)
- New Trt better for M+ (HR$_+$= 0.63)
- Interaction = 0.63/1.36 = 0.47

Prognostic and predictive;
New treatment for all?*



Quantitative interaction

- New Trt better for M– (HR_ = 0.44)
- New Trt better for M+ (HR$_+$ = 0.63)
- Interaction = 0.63/0.44 = 1.45

Interaction = HR$_+$/HR_ where HR=$\lambda_{New}/\lambda_{Std}$

(*Different considerations might apply for Standard Treatment ± New Treatment)

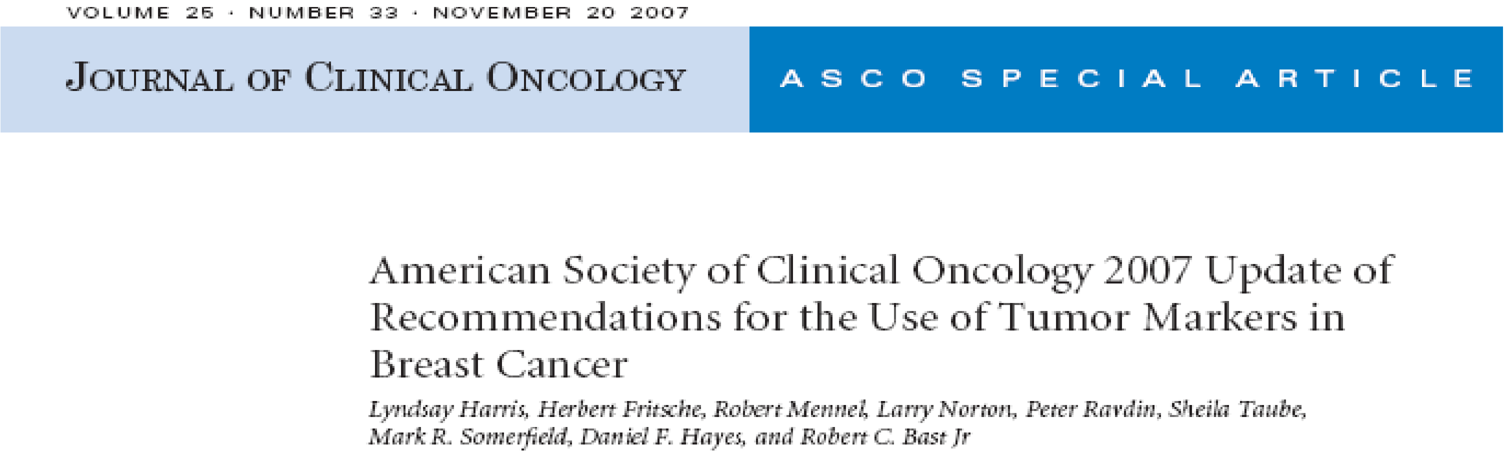
Prospective versus retrospective studies

- Prospective studies to establish clinical utility of molecular tests
  - Prognostic study design
    - Unbiased patient cohort & adjustment for standard variables
  - Predictive study designs (Freidlin et al 2010 *JNCI*; IOM Omics Report 2012)
    - Enrichment design
    - Completely randomized design
    - Biomarker-stratified design
    - Biomarker-strategy design
  - Very difficult to conduct if
    - “Take away” an established therapy
    - Prior belief in biomarker is too strong and test is already available
  - Huge and expensive

Prospective versus retrospective studies

- Retrospective studies can provide a high level of evidence if performed properly
  - Prospective-retrospective design (Simon et al 2009 *JNCI*)
    - Specimens from suitable clinical trial or well run prospective cohort study
    - Sufficient number of representative specimens
    - Analytically validated assay
    - Pre-specified analysis plan
    - Results validated in one or more similar, but separate studies
- Many retrospective studies are poorly conducted
  - No design (“convenience samples”)
  - Multiple testing and model overfitting
  - Misinterpretation
  - Deficient reporting

Profusion of retrospective studies , many of which are minimally informative



VOLUME 25 · NUMBER 33 · NOVEMBER 20 2007

JOURNAL OF CLINICAL ONCOLOGY

ASCO SPECIAL ARTICLE

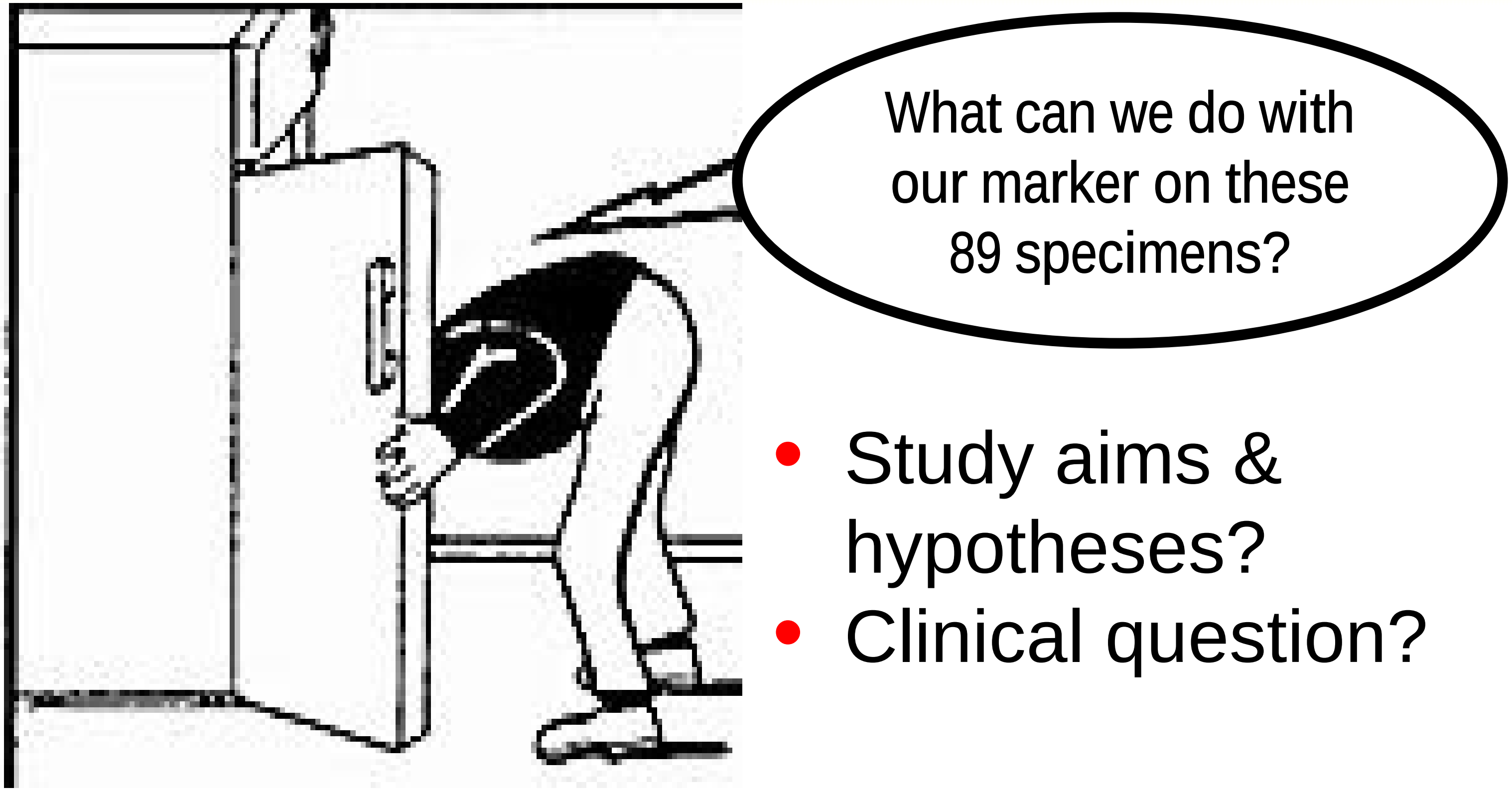
American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer

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 descriptions
of 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.

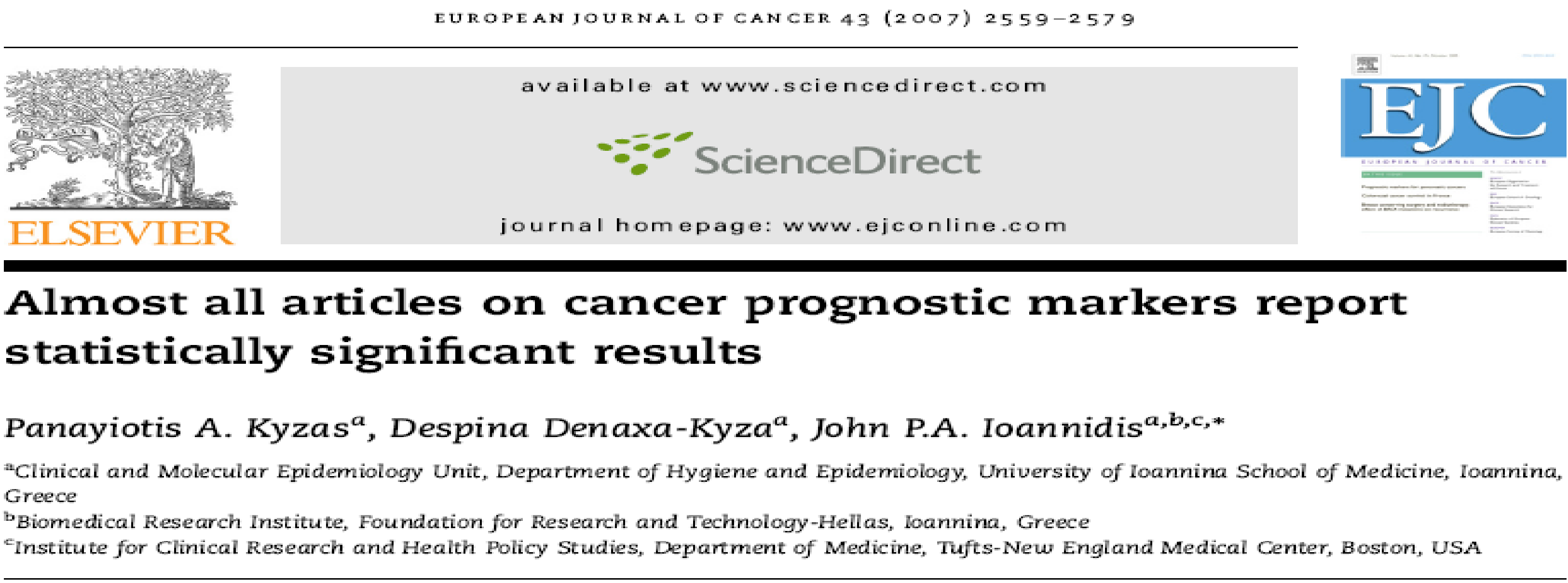
Many retrospective studies lack a design



- Study aims & hypotheses?
- Clinical question?

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

Pursuit of statistical significance . . .



EUROPEAN JOURNAL OF CANCER 43 (2007) 2559–2579

ELSEVIER

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

EJC

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

Multiple testing

# indep. tests (m) at 0.05 level	1	2	3	4	5
Probability* of ≥ 1 false positive	.05	.10	.14	.19	.23

*Prob[≥ 1 false positive] = $1-(0.95)^m$

- Multiple markers
- Multiple endpoints
- Multiple subgroups
- Multiple marker cut-points
- Multiple models

Multiple testing is particularly problematic when there is no pre-specified analysis plan and findings are selectively reported on basis of statistical significance.

Model overfitting

- Statistical model describes random error or noise instead of the true underlying relationship
  - Model is excessively complex
    - Too many parameters
    - Too many predictor variables
  - “Short fat” data
    - Many more variables than independent subjects
    - Data sparse in high-dimensional biomarker space
    - True model complex
  - Overfit model will generally have poor predictive performance on an independent data set

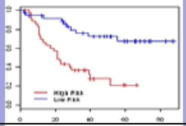
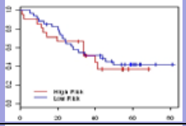
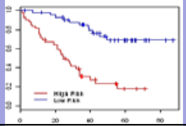
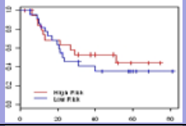
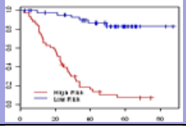
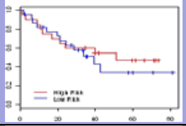
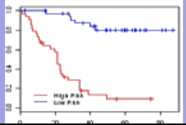
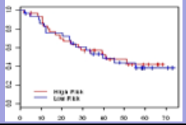
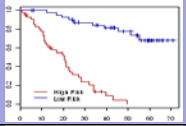
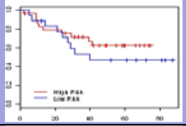
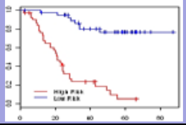
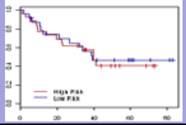
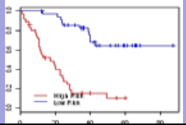
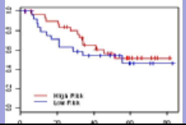
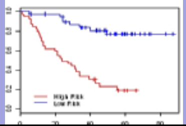
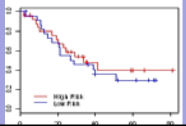
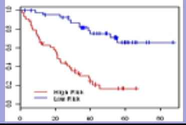
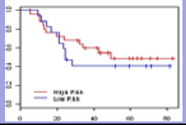
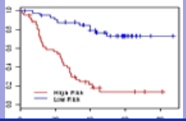
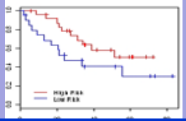
MODEL VALIDATION IS ESSENTIAL

Model validation

- RESUBSTITUTION (plug in training data) estimates of model performance are highly biased and COMPLETELY USELESS in high-dimensional data setting
- INTERNAL: Within-sample validation
  - Cross-validation
    - (Leave-one-out, split-sample, k-fold, etc.)
  - Bootstrap and other resampling methods
  - Method comparisons: Molinaro et al 2005 *Bioinformatics*
- EXTERNAL: Independent-sample validation

References: Simon et al 2003 *JNCI*;
Dupuy & Simon 2007 *JNCI*

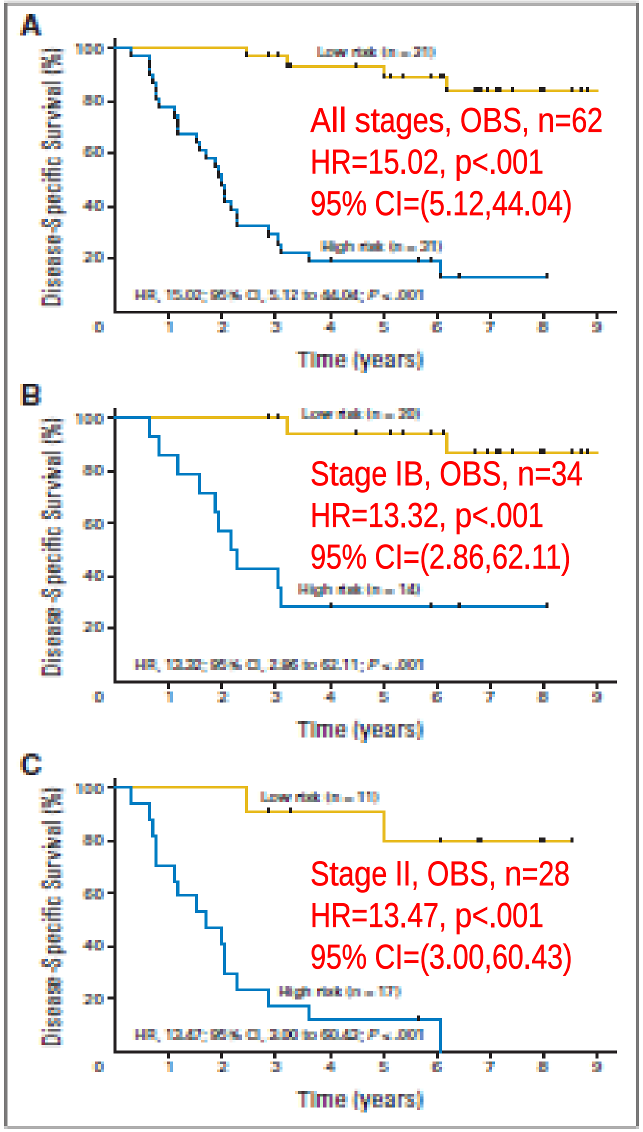
Simulation of prognostic model resubstitution method

Simulation	Training	Validation
1	 p=7.0e-05	 p=0.70
2	 p=4.2e-07	 p=0.54
3	 p=2.4e-13	 p=0.60
4	 p=1.3e-10	 p=0.89
5	 p=1.8e-13	 p=0.36
6	 p=5.5e-11	 p=0.81
7	 p=3.2e-09	 p=0.46
8	 p=1.8e-07	 p=0.61
9	 p=1.1e-07	 p=0.49
10	 p=4.3e-09	 p=0.09

(Subramanian & Simon 2010 *JNCI* – lung cancer prognostic signatures)

- Survival data on 129 patients from previous publication
- 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

Prognostic model resubstitution example



All stages, OBS, n=62
HR=15.02, p<.001
95% CI=(5.12,44.04)

Stage IB, OBS, n=34
HR=13.32, p<.001
95% CI=(2.86,62.11)

Stage II, OBS, n=28
HR=13.47, p<.001
95% CI=(3.00,60.43)

Fig 1. Disease-specific survival outcome based on the 15-gene signature in the JBR.10 training set. (A) Observation all; (B) observation stage IB; (C) observation stage II. HR, hazard ratio; ACT, adjuvant chemotherapy arm.

"A 15-gene signature separated OBS patients into high-risk and low-risk subgroups with significantly different survival (hazard ratio [HR], 15.02; 95% CI, 5.12 to 44.04; *P <.001; stage I HR*, 13.31; *P <.001; stage II HR, 13.47; P <.001)."* (*JCO 2010; 28: 4417-4424)*

Figure 1 legend:
"Disease-specific survival outcome based on the 15-gene signature in the JBR.10 training set."

Independent validations (?) of 15-gene prognostic score

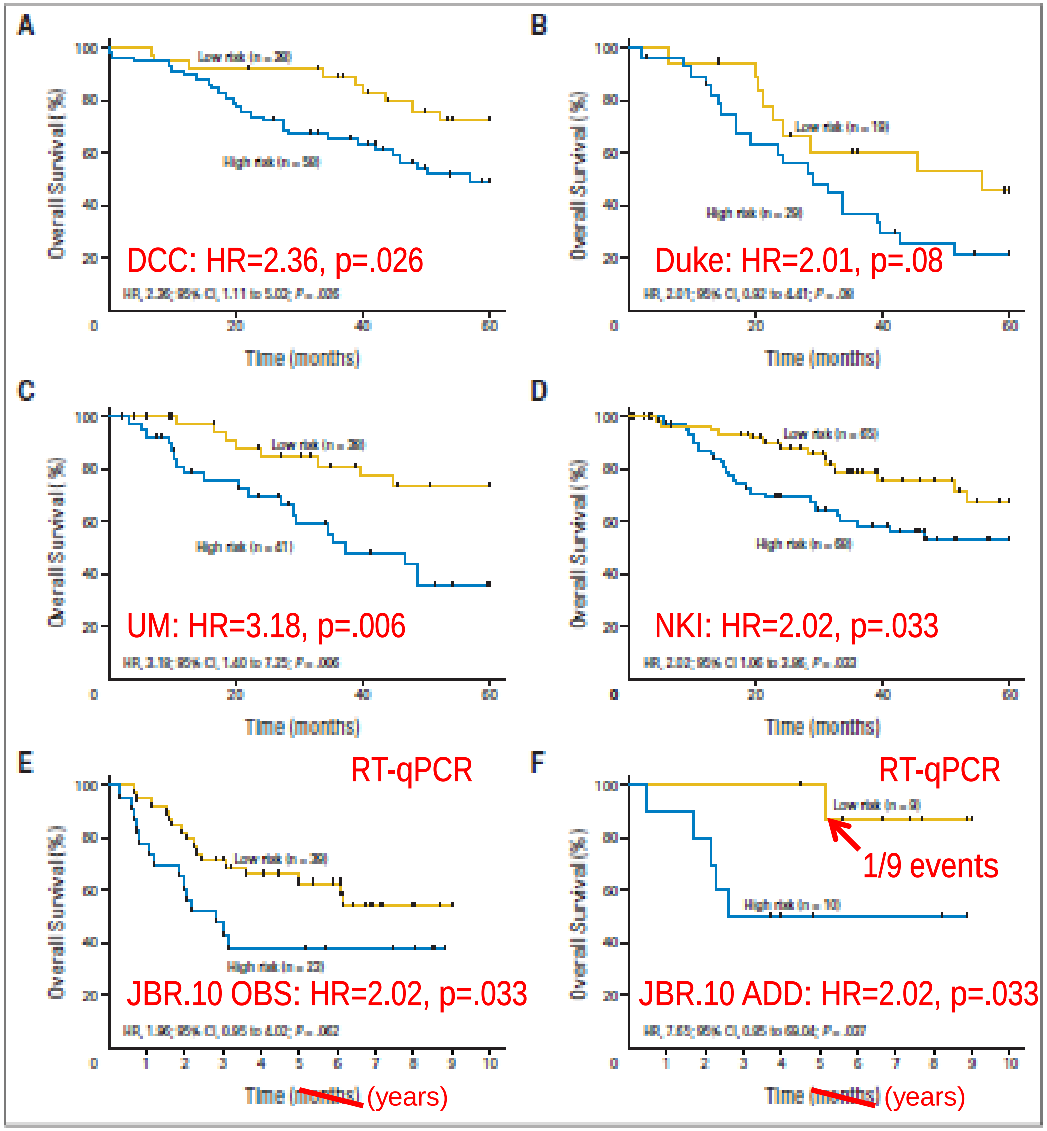


Fig 2. In silico and quantitative reverse-transcriptase polymerase chain reaction (RT-qPCR) validation of the signature in stage IB to II patients who received no adjuvant therapy. (A) Director's Challenge Consortium adenocarcinoma data set; (B) Duke University data set; (C) University of Michigan squamous cancer data set; (D) Netherlands Cancer Institute data set; (E) observation with RT-qPCR; (F) observation with RT-qPCR with additional samples. HR, unadjusted hazard ratio.

"The prognostic effect was verified in the same 62 OBS patients where gene expression was assessed by qPCR. Furthermore, it was validated consistently in four separate microarray data sets (total 356 stage IB to II patients without adjuvant treatment) and additional JBR.10 OBS patients by qPCR (n=19)."

What happened to HR=15.02? Different endpoint?

Is this still clinically useful?

Assessment of predictive tests: Resubstitution pitfalls again

Is resubstitution acceptable when model was fit using the control (OBS) arm only? NO! (Fig. 3, *JCO 2010; 28: 4417-4424*)

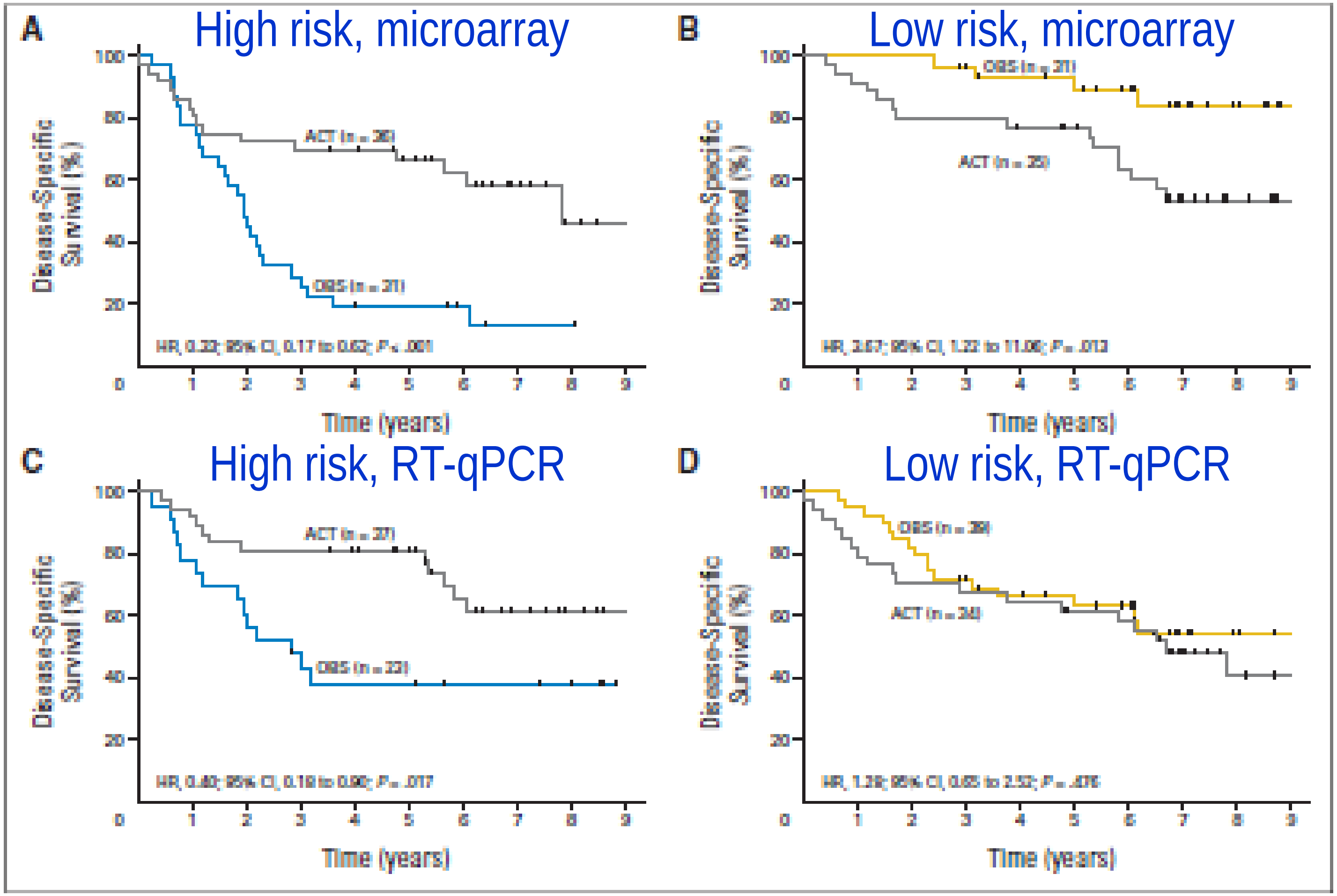


Fig 3. Predictive effect of the signature to adjuvant chemotherapy. Only high-risk group benefits from adjuvant chemotherapy. (A) High risk (microarray); (B) low risk (microarray); (C) high risk quantitative reverse-transcriptase polymerase chain reaction (RT-qPCR); (D) low risk RT-qPCR.

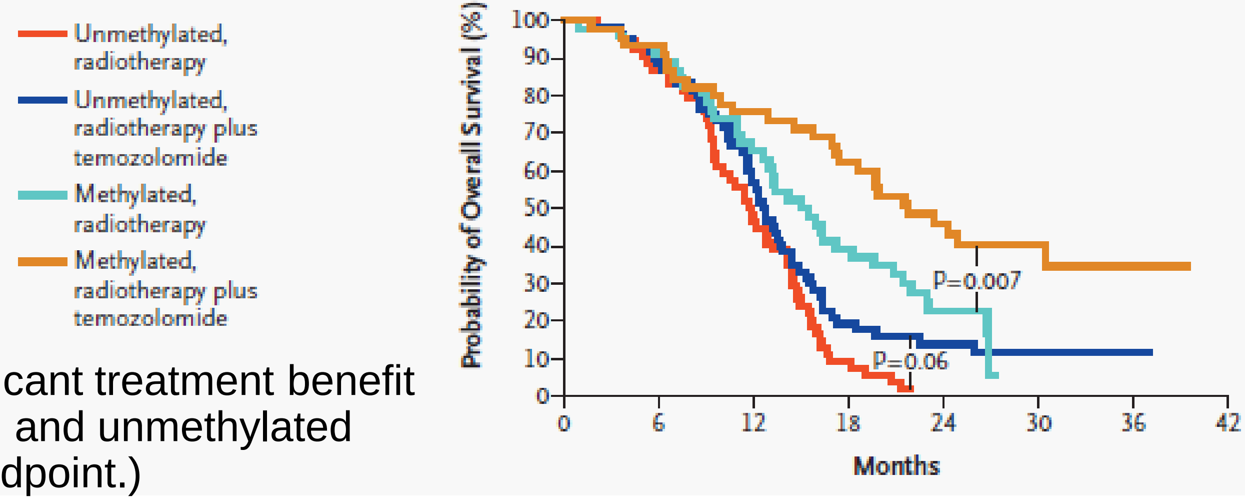
Assessment of predictive tests: Power pitfalls

- Randomized clinical trials adequately powered to detect treatment effects are often not sufficiently powered to establish predictive marker effects
- Non-significance of treatment effect in a “marker negative” subgroup is often misinterpreted as no treatment effect

Assessment of predictive tests: Power pitfalls

CONCLUSION: "Patients with glioblastoma containing a methylated *MGMT* promoter benefited from temozolomide, whereas those who did not have a methylated *MGMT* promoter did not have such a benefit."

(NEJM 2005; 352: 997-1003)



(Statistically significant treatment benefit in both methylated and unmethylated groups for PFS endpoint.)

Overall Survival (OS)	Hazard ratio (95% CI)	Median OS (months)	2-yr OS (%)	P-value
MGMT Methylated				
RT (n=46)	1.00	15.3 (13.0-20.9)	22.7 (10.3-35.1)	
RT+TMZ (n=46)	0.51 (0.31-0.84)	21.7 (17.4-30.4)	46.0 (31.2-60.8)	0.007
MGMT Unmethylated				
RT (n=54)	1.00	11.8 (9.7-14.1)	< 2	
RT+TMZ (n=60)	0.69 (0.47-1.02)	12.7 (11.6-14.4)	13.8 (4.8-22.7)	0.06

Assessment of predictive tests: Power pitfalls

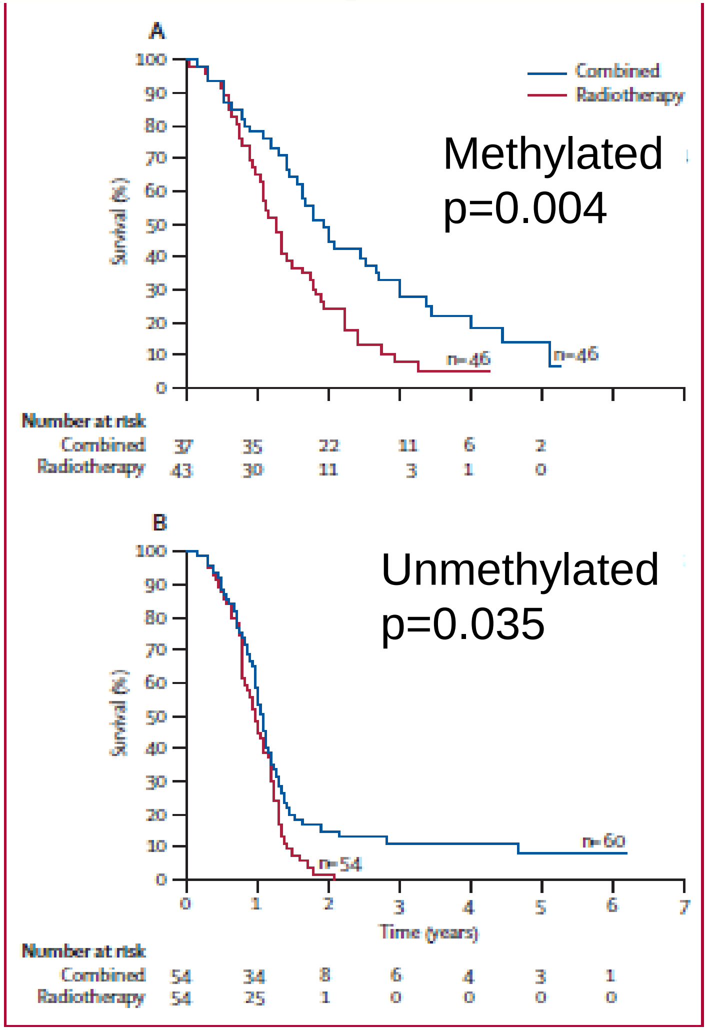


Figure 4: Kaplan-Meier estimates of overall survival by MGMT status
Patients with methylated *MGMT* (A). Patients with unmethylated *MGMT* (B).

(Salvage therapies, including TMZ, confound OS endpoint.)

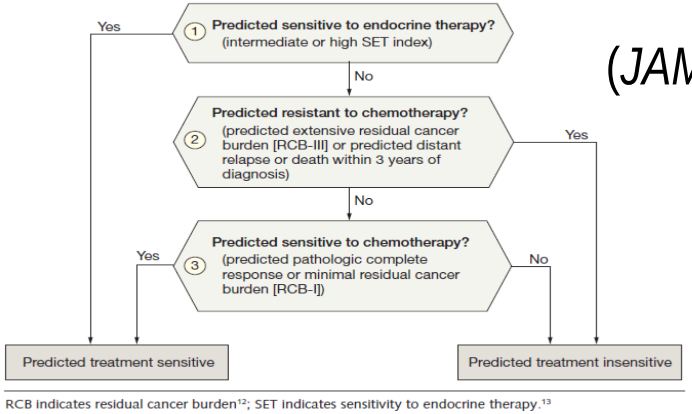
With follow-up to 5 years, the OS difference became significant in favor of RT+TMZ even in the unmethylated *MGMT* group (not adjusted for testing in 2 subgroups).
(*Lancet Oncol* 2009; 10: 459-466)

	Hazard ratio (95% CI)	5-yr OS (%)
MGMT Methylated		
RT	1.0	5.2 (1.0-15.0)
RT + TMZ	0.3 (0.2-0.4)	13.8 (4.5-28.2)
MGMT Unmethylated		
RT	1.0	0
RT+TMZ	0.6 (0.4-0.8)	8.3 (2.7-18.0)

Assessment of predictive tests: Pitfalls of non-randomized comparisons

Figure 1. Genomic Decision Algorithm to Predict Sensitivity of Invasive Breast Cancer to Adjuvant Chemotherapy (CT) or Chemoendocrine Therapy (CT+ HT)

(*JAMA* 2011; 305: 1873-1881)



1 Predicted sensitive to endocrine therapy? (intermediate or high SET index)

Yes → Predicted treatment sensitive

No ↓

2 Predicted resistant to chemotherapy? (predicted extensive residual cancer burden [RCB-III] or predicted distant relapse or death within 3 years of diagnosis)

Yes → Predicted treatment insensitive

No ↓

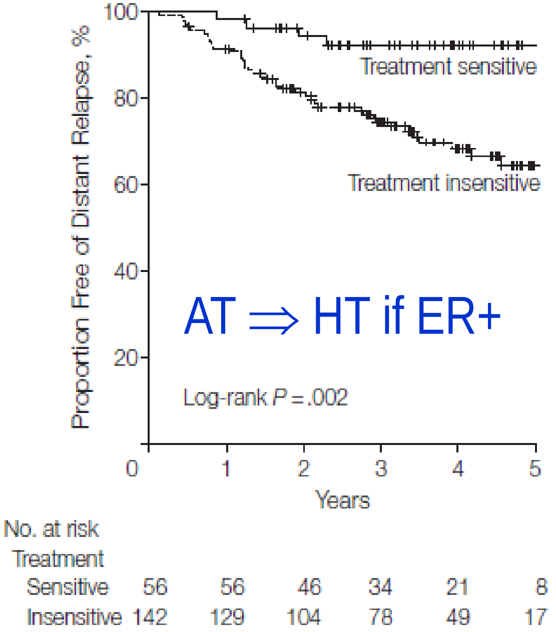
3 Predicted sensitive to chemotherapy? (predicted pathologic complete response or minimal residual cancer burden [RCB-I])

Yes → Predicted treatment sensitive

No → Predicted treatment insensitive

RCB indicates residual cancer burden[12]; SET indicates sensitivity to endocrine therapy.[13]

Figure 2.
Validation Cohort #1

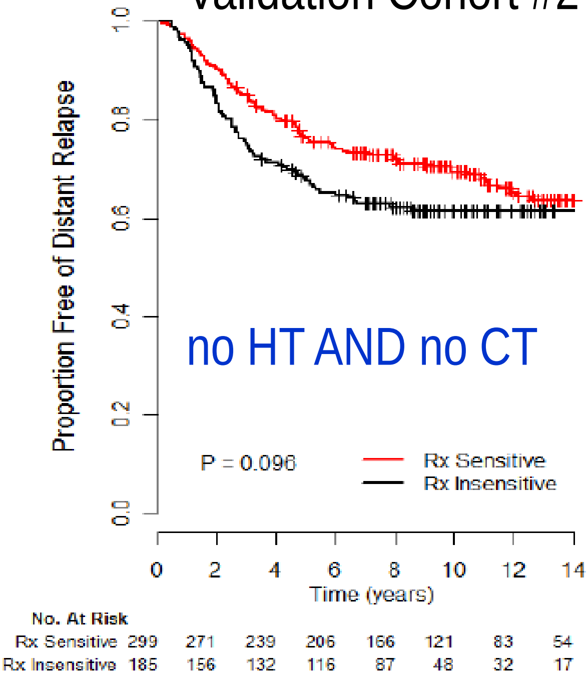


AT ⇒ HT if ER+

Claim: Test is predictive and not prognostic
P=.002 (Fig 2) vs.
P=.096 (eFig 6A)

A = anthracycline
T = taxane
HT = hormonal therapy

eFigure 6A.
Validation Cohort #2



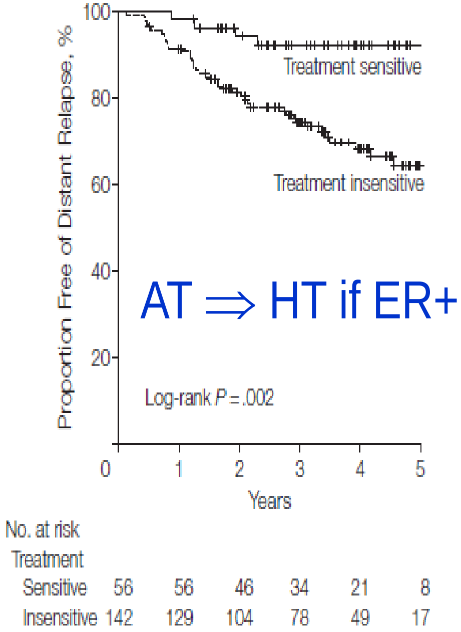
no HT AND no CT

Assessment of predictive tests: Pitfalls of non-randomized comparisons

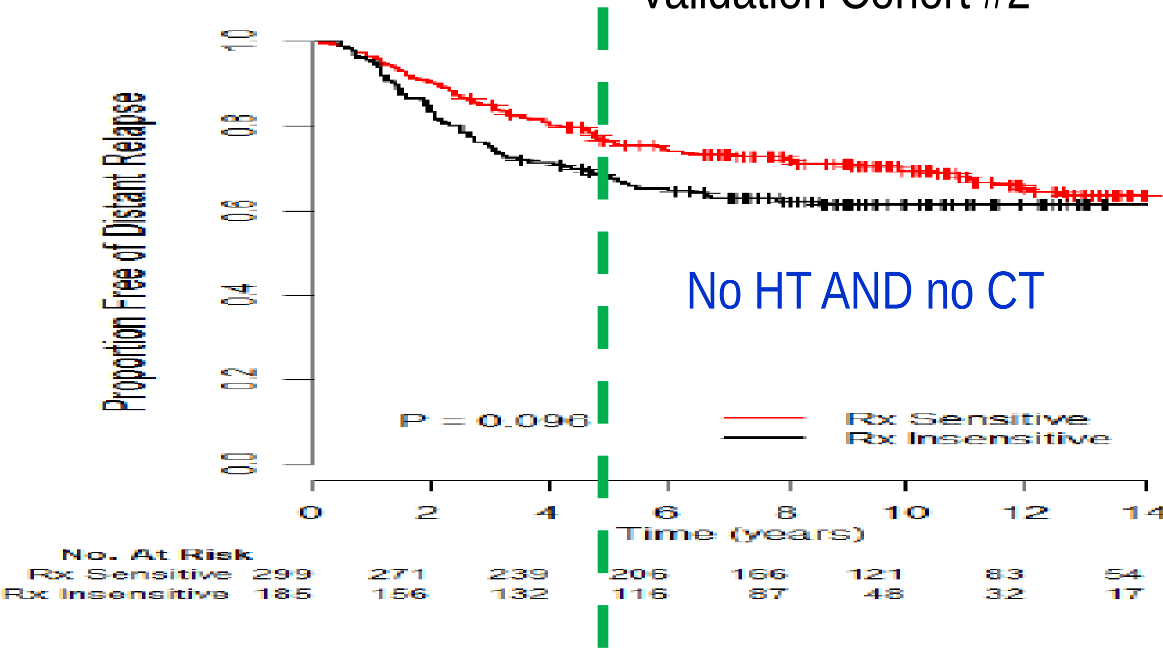
Figure 2.
Validation Cohort #1

(*JAMA* 2011; 305: 1873-1881)

eFigure 6A.
Validation Cohort #2



AT ⇒ HT if ER+



No HT AND no CT

Differences between validation cohorts

Cohort 1	Cohort 2
35% N–, 65% N+ (worse prognosis)	100% N– (better prognosis)
62% ER+	71% ER+
All ER+ receive endocrine therapy	No endocrine therapy
All receive taxane	No taxane therapy
Follow-up ends at 5 yrs	Curves merge around 14 yrs

Press Release: Ready to Launch Tests for HT, CT Response in Breast Cancer

NEW YORK – Following the results of a study suggesting that its genomic test may have use in predicting chemotherapy response in patients with breast cancer, <Company> said that a launch of the test, as well as another for predicting endocrine therapy response, is in the works.

. . .

In the study, published in the **May 11 issue of the *Journal of the American Medical Association***, the authors said that patients who were predicted to be sensitive to taxane-anthracycline chemotherapy had a 56 percent probability of "excellent pathologic response" and distant relapse-free survival of 92 percent, as well as an absolute risk reduction of 18 percent.

. . .

Based on those results, <Company> is in the process of validating the test for **launch in a CLIA format and is now seeking a commercialization partner. And during the second half of this year, the company anticipates it will embark on a strategy to receive clearance from the US Food and Drug Administration for the test**.

Summary recommendations

- Earlier and more intense focus on clinical utility
  - Educate about proper interpretation
- Rigor in test development process and study design
  - Meaningful well-designed studies
  - Proper statistical analysis
  - Independent external validation
  - Inter-disciplinary expertise
- Biomarker study registry (Andre et al 2011, *Nat Rev Clin Oncol*) (http://win.biomarkerregistry.org)
  - Aid in identifying relevant biomarker studies for overviews and meta-analyses
  - Submission of study protocols (pre-specified analysis plans)
  - Help reduce non-publication bias and selective reporting

Summary recommendations (cont.)

- Complete and transparent reporting
  - REMARK guidelines
    - McShane et al 2005 *J Natl Cancer Inst*
    - Altman et al 2012, *BMC Med* and *PLoS Med* – E&E
  - EQUATOR Network – collection of reporting guidelines for health research studies (www.equator-network.org)
  - BRISQ – reporting details of biospecimen collection, handling, storage (Moore et al 2011 *Cancer Cytopathol)*
  - McShane & Hayes (*JCO*, in press)
- Expanded access to *useful* specimens
  - Well-annotated with clinico-pathologic data, treatment, and clinical outcome
  - Alternative sources (trial specimens optimal but limited)
- Alignment of good science, regulation, and payment