Assessing readiness of an omics signature for use in a clinical trial

Improving Cancer Diagnosis and Care: The Clinical Application of Computational Methods in Precision Oncology

National Academies of Sciences, Engineering, and Medicine
Washington, DC

Lisa M McShane, PhD
Biometric Research Program, Division of Cancer Treatment and Diagnosis

NIH NATIONAL CANCER INSTITUTE

October 29-30, 2018

Disclaimers

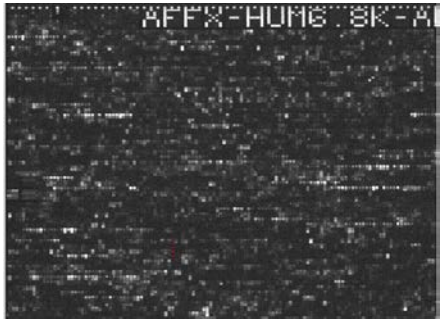
- The views expressed represent my own and do not necessarily represent views or policies of the U.S. National Cancer Institute.
- Examples I cite are all true, but in some cases I have concealed details to protect identities.

My perspective

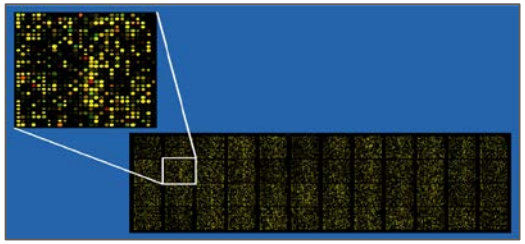
- Statistical/scientific reviewer of NCI-sponsored clinical trials and studies for development and validation of biomarker-and omics-based tests
- Scientific Advisory Board (*Science Translational Medicine*) and Editorial Board (*BMC Medicine*)
- Statistical reviewer for numerous biomedical journals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

Omics

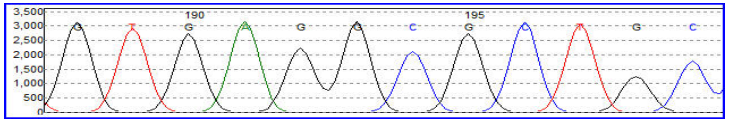
"A term encompassing multiple molecular disciplines, which involve the characterization of global sets of biological molecules such as DNAs, RNAs, proteins, and metabolites."



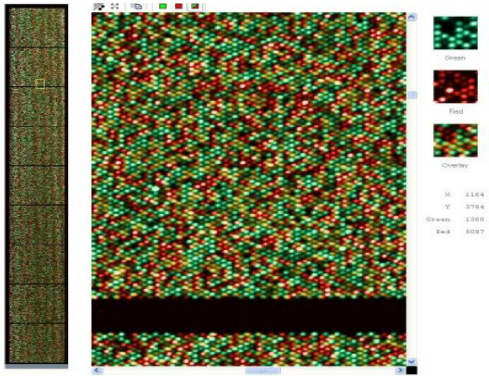
Affymetrix expression GeneChip



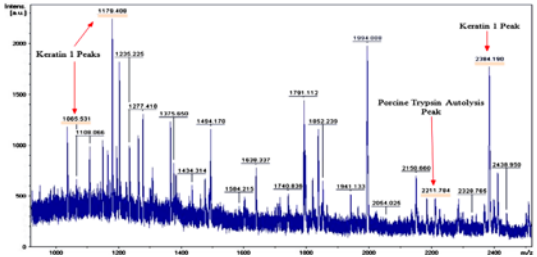
cDNA expression microarray



Mutation sequence surveyor trace



Illumina SNP bead array



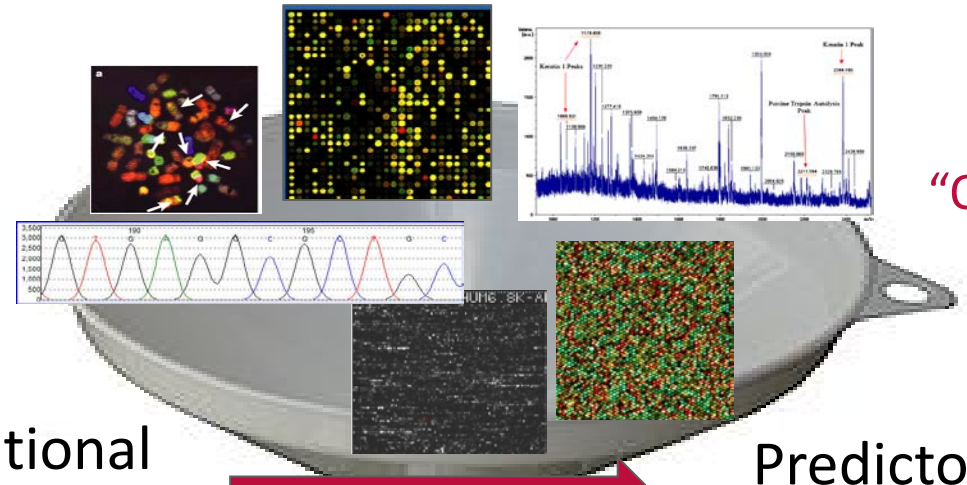
MALDI-TOF proteomic spectrum

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

Translation from omics discoveries to clinically useful omics-based tests

Discovery

High-throughput omics assays

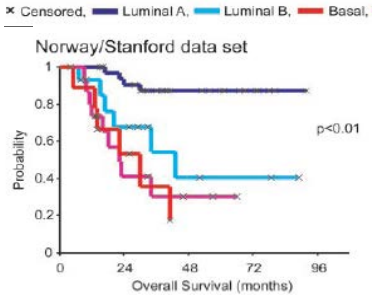


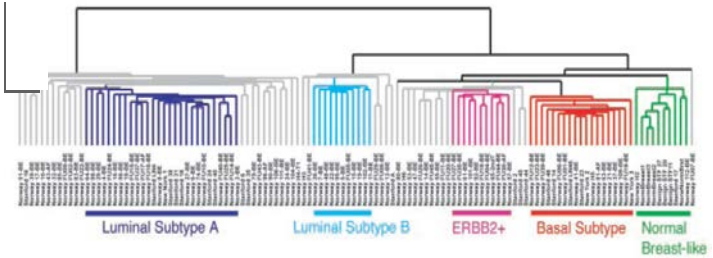
"Omics predictor"

Computational models

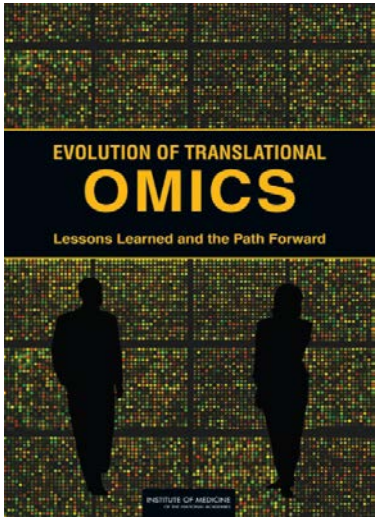
Predictors, classifiers, risk scores

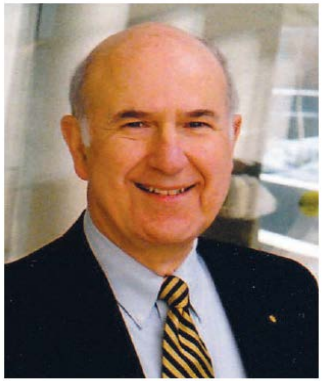
Clinical Utility?





Institute of Medicine reviews field of translational omics





"There are a lot of lessons here that surely apply to other places."

—GILBERT S. OMENN, UNIVERSITY OF MICHIGAN, ANN ARBOR

"There are a lot of lessons here that surely apply to other places."

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

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

McShane et al. Nature 2013;502:317-320 (checklist)
McShane et al. BMC Medicine 2013;11:220 (explanation & elaboration)

5 DOMAINS

- ✓Specimens
- ✓Assays

} Getting all of the details right is hard work; important but often ignored in discovery stage

- ✓Model development, specification & *preliminary* performance evaluation

} Too many researchers still fall into traps of high-dimensional data analysis

- Clinical trial design
- Ethical, legal, and regulatory issues

} No time to discuss today

Domain 1: Specimens

- Collection, processing & storage (pre-analytic factors)
- Specimen quality screening
- Minimum required amount
- Enrichment (e.g., micro- or macro-dissect for tumor cells)
- Feasibility of collecting needed specimens
  - Specimen collection (e.g., biopsy) is safe
  - Achievable in standard clinical settings
  - Percent useable for assays is sufficiently high

Domain 1: Specimen pre-analytic requirements and quality monitoring plans

- Example 1: Only 20% of first 100 specimens collected in a diagnostics study were of adequate quality for immediate assay due to failure to freeze immediately upon collection
- Example 2: Only 50% of frozen samples collected on a trial were fit for assay upon thawing a few years later

Moore et al. Biospecimen Reporting for Improved Study Quality (BRISQ). Cancer Cytopathology 2011;119:92-101

Domain 2: Assay considerations

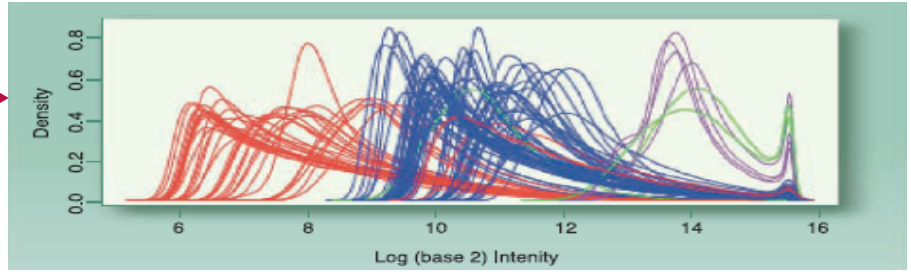
- Lock down standard operating procedures (SOPs)
- Impact of changes in assay procedures
- Quality monitoring (controls, standards)
  - Equipment malfunction, bad reagent lots, re-calibration needed
- Quality criteria to accept/reject assay values
  - Bad specimens, batch effects
- Analytical performance evaluation (sensitivity, specificity, bias, accuracy, precision, reproducibility)
  - Becker R, Analytical validation of in vitro diagnostic tests. In *Design and Analysis of Clinical Trials for Predictive Medicine*, Matsui, Buyse, Simon (eds.), Chapman and Hall/CRC, 2015, pp. 33-50.
  - Jennings et al, *Arch Pathol Lab Med* 2009;133: 743–755
- Turnaround time

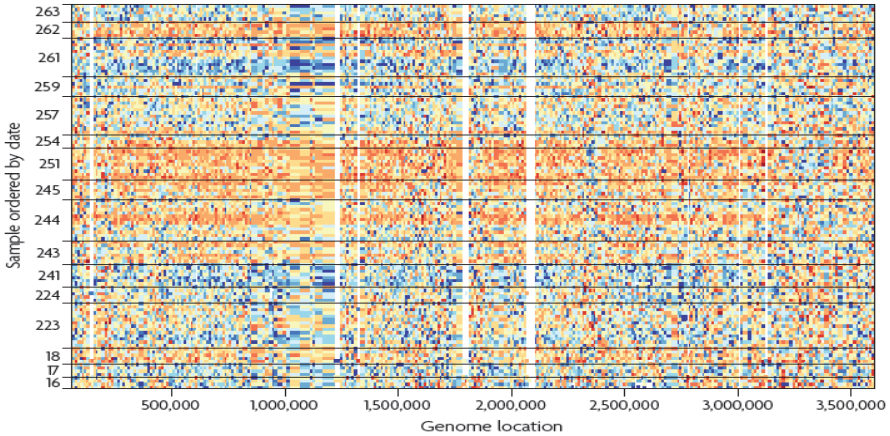
Domain 2: Assay artifacts & batch effects

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

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

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





Batch effects for 2nd generation sequence data (std. coverage data). Same facility & platform.

Horizontal lines divide by date

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

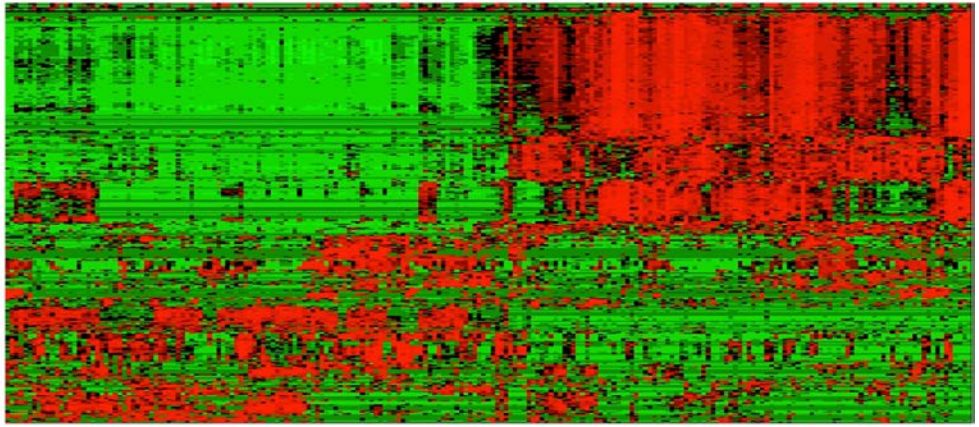
BATCH EFFECTS ARE ESPECIALLY PROBLEMATIC IF CONFOUNDED WITH KEY EXPERIMENTAL FACTORS OR ENDPOINTS.

Domain 2: Assay artifacts & batch effects

- Impact of changes in assay procedures, reagents, equipment, or technician during predictor development

  Dramatic effect of change in RNA extraction procedure & reagents on tumor gene expression microarray profiles

Extraction method 1 ¦ Extraction method 2



215 tumor samples

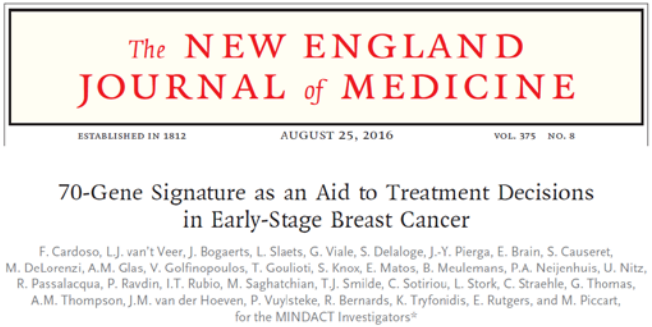
116 genes included in a genomic predictor of treatment response

(Shown with permission from an NIH grantee)

Domain 2: Assess impact of changes in assay reagents or procedures before implementing

MINDACT Trial

Cardoso F et al., *N Engl J Med* 2016;375:717-729



"A **change in the RNA-extraction solution** that was used in the calculation of the 70-gene signature (a change that was not communicated by the manufacturer) **caused a temporary shift in the risk calculation** from May 24, 2009, to January 30, 2010, at which time the issue was discovered and rectified . . .

Because of this shift, **162 patients who had been identified as being at high genomic risk were subsequently identified as being at low genomic risk** with the use of the correct . . .

The clinical effect of this risk revision was that an **additional 28 patients received chemotherapy before the results were corrected, although no patient was undertreated**."

Domain 3: Model development & preliminary evaluation

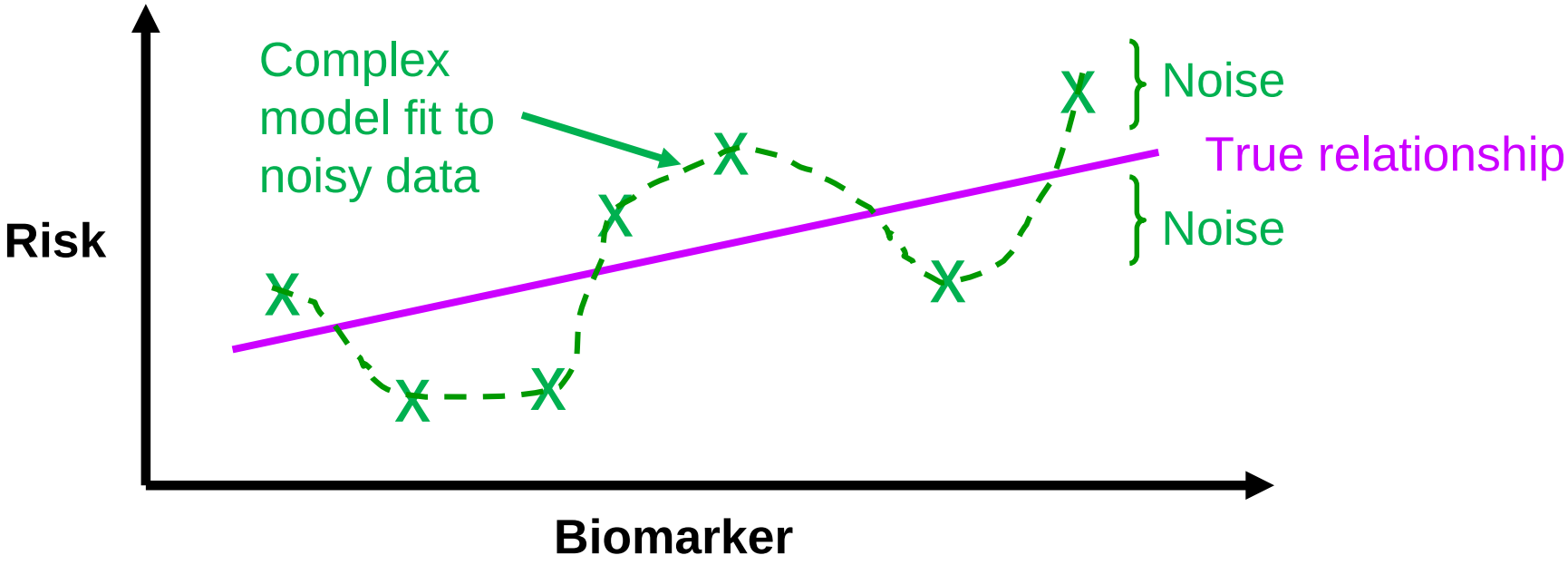
- Quality of data (clinical & omics) used to develop and validate predictor models (might not be “clinical trials grade” data)
- Appropriate statistical approaches for model development and performance assessment
- Appropriate “validation”
  - Define clinical context and use
    - Patient population
    - Clinical use - prognostic, predictive, etc.
  - “Locked down” test
  - Pre-specified evaluation criteria (not just a significant p-value)

Domain 3: Model development & preliminary evaluation common pitfalls

- A statistical model is **OVERFIT** when it describes random error or noise instead of the true underlying relationship
  - Excessively complex (too many parameters or predictor variables)
  - Will have poor predictive performance on independent data set
  - Naively fit omics predictors will always be overfit
- **RE-SUBSTITUTION** is the naïve evaluation of model performance by "plugging in" same data used to build it
  - Other more subtle forms of re-substitution (combining training & test, with covariates, comparative, partial)

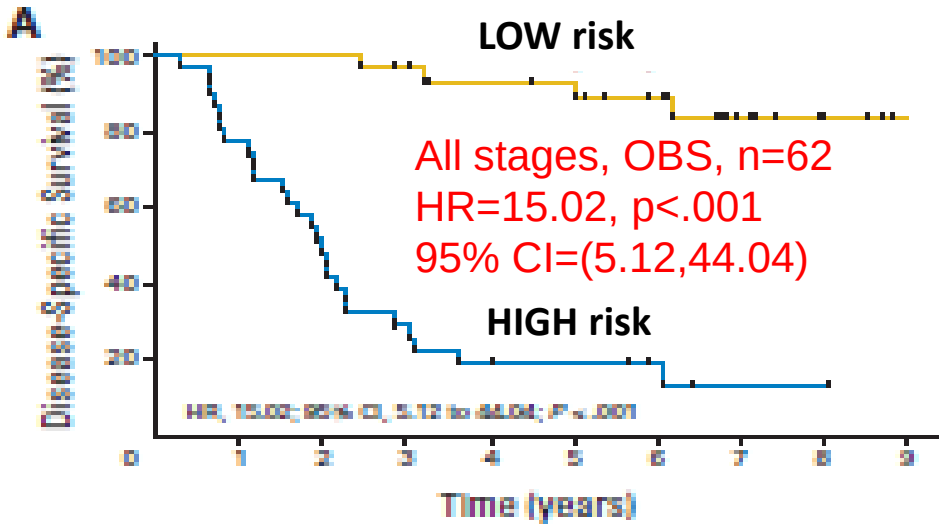
(J Biopharm Statistics 2016;26(6):1098-1110)

Domain 3: Model over-fitting



- Evaluation of a model's fit by data re-substitution will suggest fit is perfect
- In high dimensions (e.g., omics data), naively fit models are almost always over-fit and such models will rarely validate on an independent data set

Domain 3: Avoid "re-substitution" pitfall



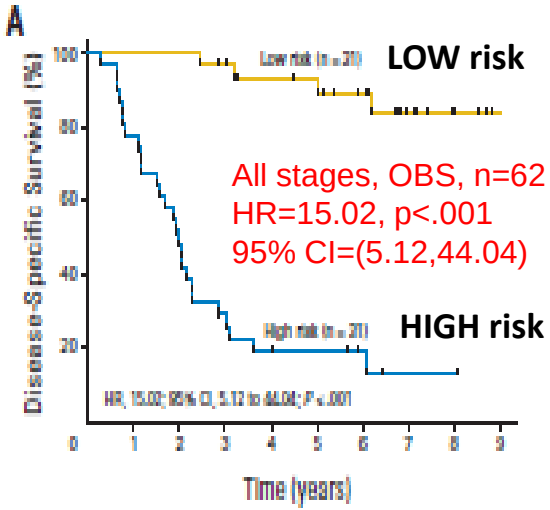
(J Clin Oncol 2010;28:4417-4424)

"A 15-gene signature *[for lung cancer]* separated OBS patients *[no chemotherapy after surgery]* 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."*

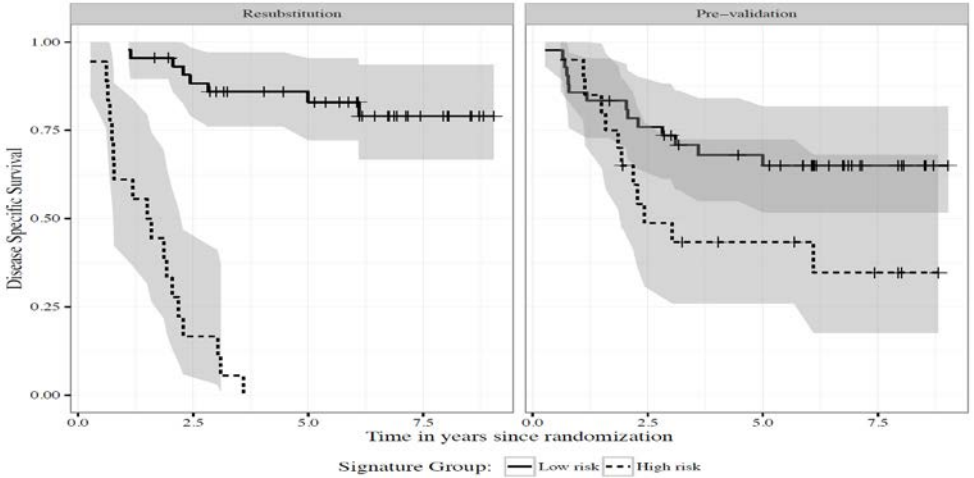
RE-SUBSTITUTION!

If this large separation in survival curves was real, the signature would have clinical utility. Patients designated as low risk could confidently avoid toxic chemotherapy.

Domain 3: Internal validation (e.g., cross-validation) to correct for re-substitution bias



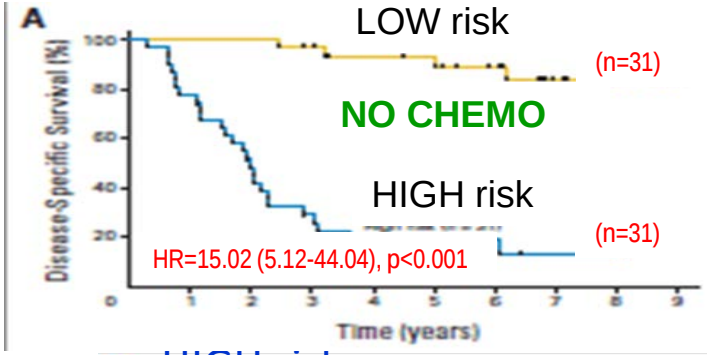
Original Kaplan-Meier curves (DSS) showing prognostic ability of 15-gene signature in OBS arm, using re-substitution
(J Clin Oncol 2010;28:4417-4424)



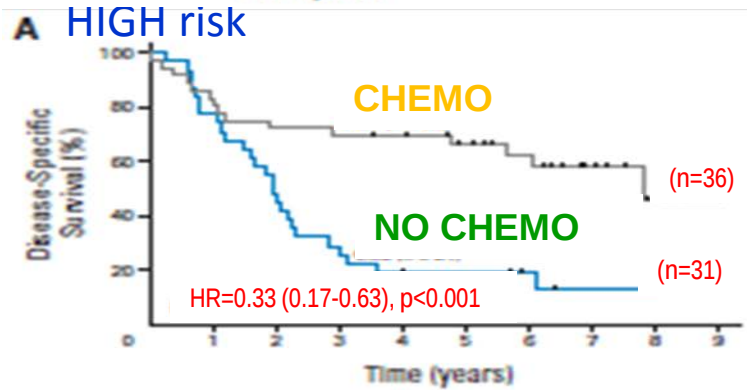
RE-SUBSTITUTION CROSS-VALIDATION

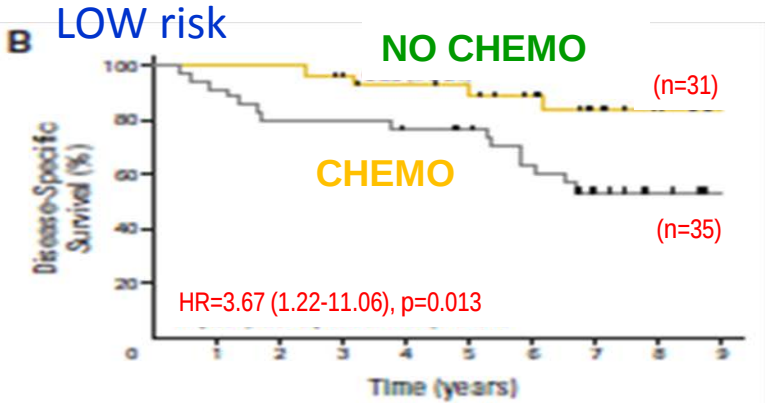
Reproduced (approx.) Kaplan-Meier curves (DSS) showing prognostic ability of 15-gene signature in OBS arm, using re-substitution (LEFT) and cross-validation (RIGHT)
(J Biopharm Statistics 2016;26(6):1098-1110)

Domain 3: Avoid *comparisons* with re-substitution estimates



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

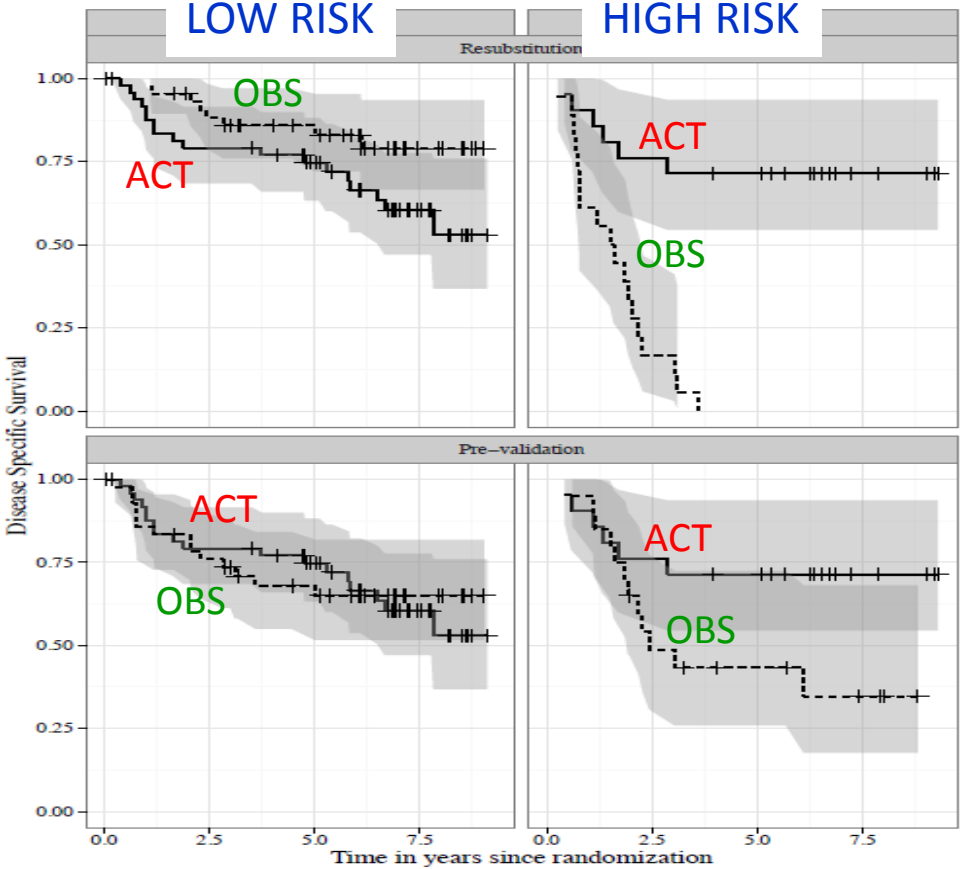




Does the genomic predictor identify groups of patients who benefit differently from adjuvant chemotherapy? **Can't conclude anything.**

Domain 3: Internal validation (e.g., resampling)

Use of cross-validation to correct for re-substitution bias



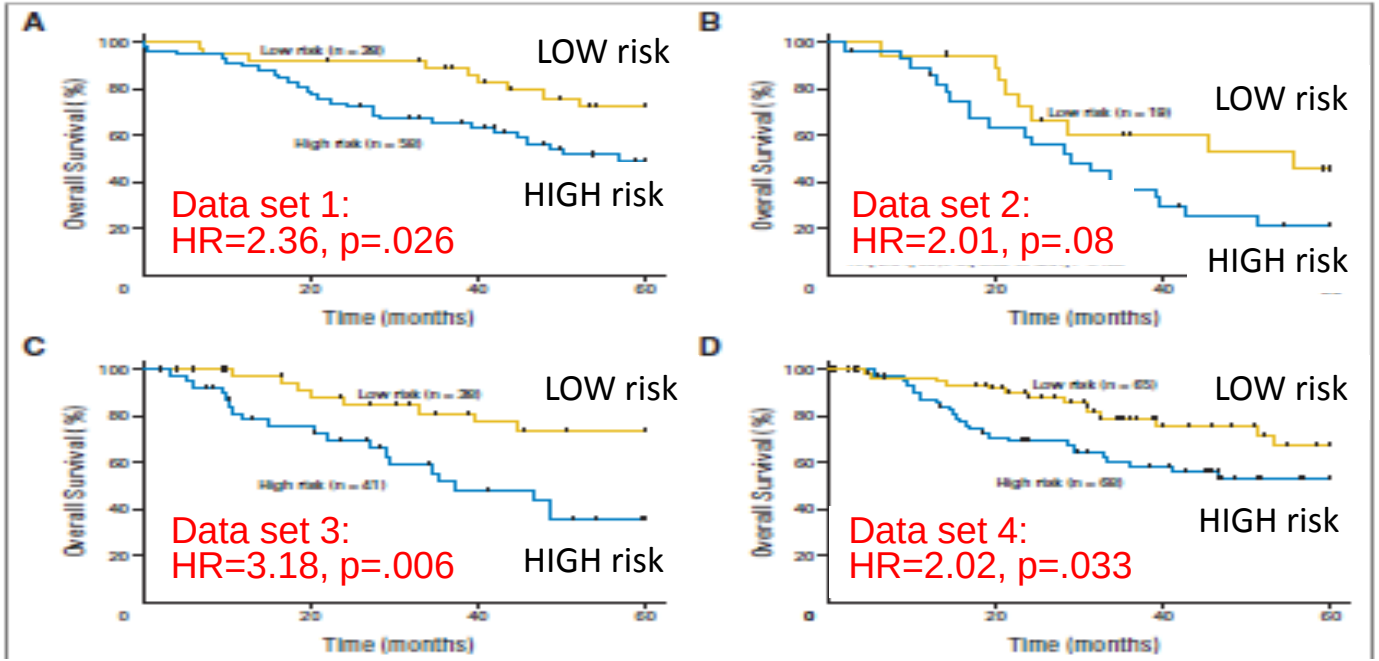
RE-SUBSTITUTION

CROSS-VALIDATION

Reproduced (approx.) Kaplan-Meier curves (DSS) suggesting ***predictive ability*** (for treatment-selection) of 15-gene signature using re-substitution (TOP) and cross-validation (BOTTOM)

(J Biopharm Statistics 2016;26(6):1098-1110)

Domain 3: What do we mean by validation?



(J Clin Oncol 2010;28:4417-4424)

"... prognostic effect *[of 15-gene signature]* was **VALIDATED** consistently in four separate microarray data sets (total 356 stage IB to II patients without adjuvant treatment)."

Endpoint: Disease-specific survival (DSS) → Overall survival (OS)
Timescale: 0 to 9 years → 0 to 60 months (5 years)
HR: 15.02→ ≈ 2-3 5-yr DSS ≈ 90% → 5-yr OS < 80%
Mixture of disease stages? Adjustment for standard covariates?

Domain 3: What do we mean by validation?

- "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma"
- Original article: "We elucidated a neoantigen landscape that is specifically present in tumors with a strong response to CTLA-4 blockade. We **validated** this signature in a second set of 39 patients with melanoma who were treated with anti–CTLA-4 antibodies."
- Correction: "Some readers were confused by our incomplete description of part of the data analysis and our use of the term "validation set . . . **our use of "validation set" was not appropriate** in the context of the search for a neoantigen signature, since information from both data sets was used to derive the results.. . we **did not use "validation set" in the conventional way. In contrast to a formal biomarker analysis, our study design focused on defining a recurrent genetic footprint that occurred in a nonrandom fashion.** (Several other corrections made as well.)

N Engl J Med 2015;373(20):1989-1989

Domain 3: Requirements for a rigorous validation of a predictor

- Predictor must be completely **LOCKED DOWN** and must be a **PRE-SPECIFIED PERFORMANCE METRIC**. The lockdown includes *all steps* in the data pre-processing and prediction algorithm (including computer code).
- Ideally, **INDEPENDENT VALIDATION DATA** 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 according to the **PRE-SPECIFIED ASSAY** protocol (including quality rejection criteria).

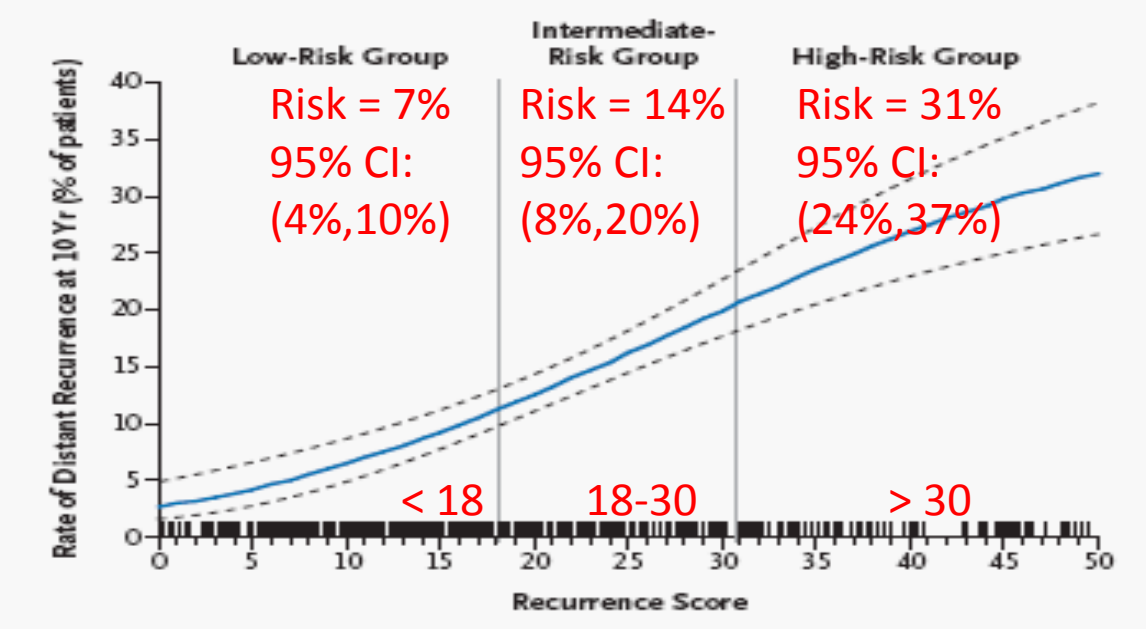
(cont →)

Domain 3: Requirements for rigorous validation of a predictor (cont.)

- Individuals who developed the predictor must remain completely **BLINDED** to the validation data.
- The validation **DATA SHOULD NOT BE CHANGED** and **DATA VALUES SHOULD NOT BE SELECTIVELY ELIMINATED** after observing the performance of the predictor.
- **PREDICTOR SHOULD NOT BE ADJUSTED** (including cut-points) 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.

SUCCESS STORY (part 1): Rigorous validation of Oncotype DX recurrence score (RS) prognostic ability

Prospective-Retrospective Validation of RS on NSABP B-14 Tamoxifen Arm (tamoxifen treated ER+ breast cancer)

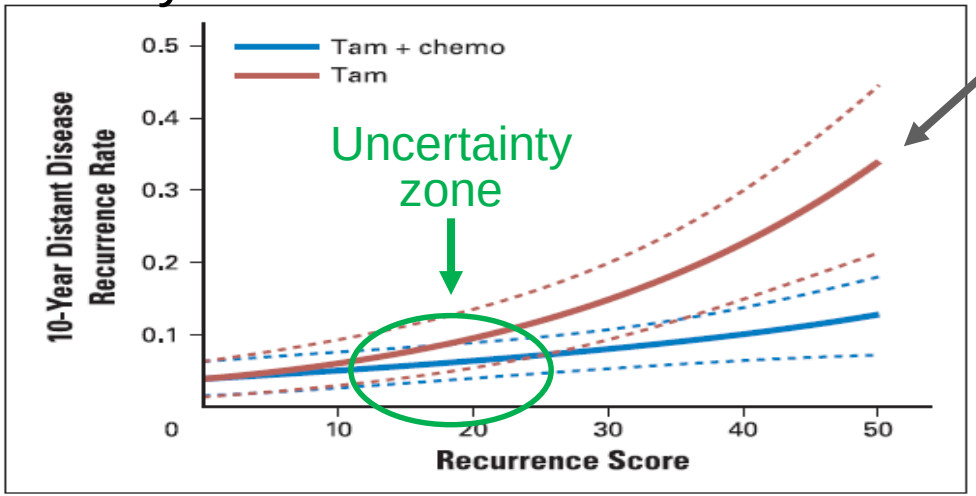


(Figure 4 from Paik et al., N Engl J Med 2004;351:2817-26)

- 21-gene RT-PCR based gene expression assay
- FFPE tissues
- Locked assay and RS algorithm
- RS assignments blinded to outcome data
- "Honest broker" linked final RS to clinical outcome data

SUCCESS STORY (part 2): Is Oncotype DX predictive for chemotherapy benefit?

Preliminary Evaluation of Predictive Ability of RS on NSABP B-20



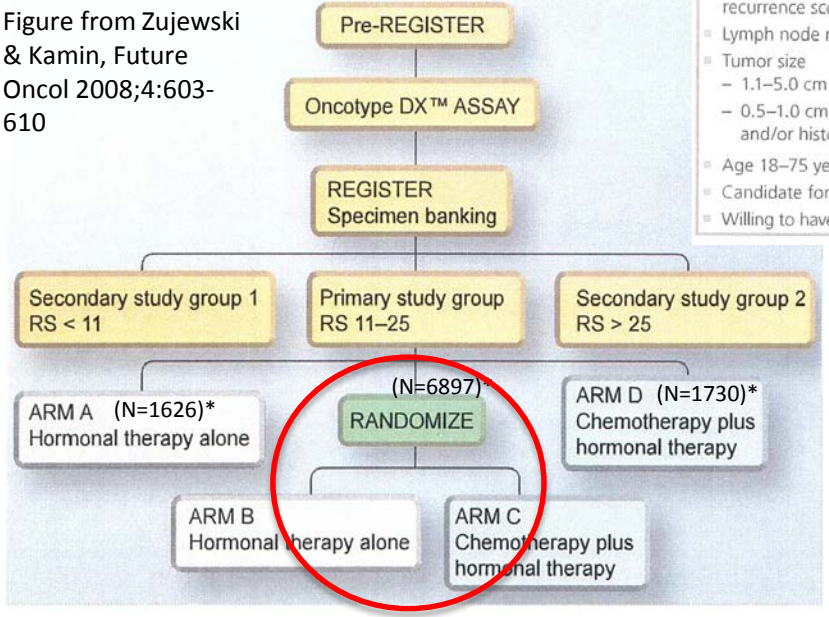
- Possible bias due to use of Tam arm in RS development?
- Few events in the intermediate zone (imprecise estimate of chemo benefit)
- Impact of newer endocrine therapies (e.g., AIs)

This study alone was considered insufficient to establish clinical utility of Oncotype DX for ***therapy selection (predictive ability)***.

Paik et al., *J Clin Oncol* 2006;24:3726-3734 (Figure 4)

TAILORx trial design

Figure from Zujewski & Kamin, Future Oncol 2008;4:603-610



Box 1. Patient eligibility criteria.

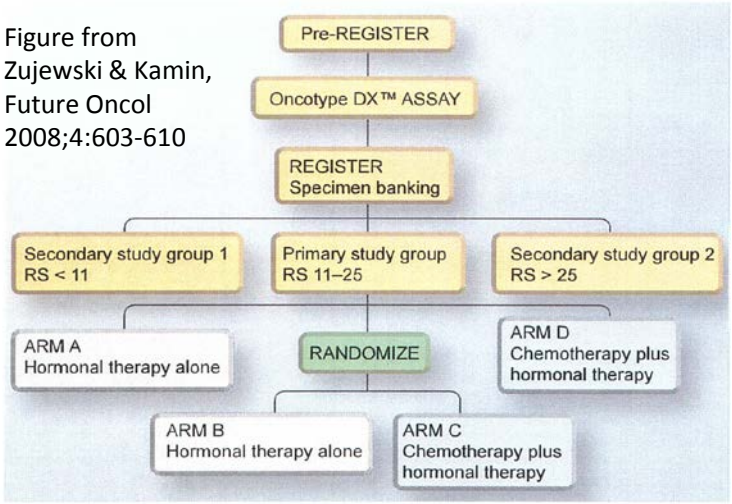
- Estrogen-receptor- and/or progesterone-positive breast cancer
- Her2/neu negative (re: most Her2-positive disease associated with high recurrence score)
- Lymph node negative (by sentinel node or axillary dissection)
- Tumor size
  - 1.1–5.0 cm
  - 0.5–1.0 cm plus unfavorable histologic features (intermediate or poor nuclear and/or histologic grade, or lymphovascular invasion)
- Age 18–75 years
- Candidate for systemic chemotherapy
- Willing to have treatment assigned or to be randomized based upon Oncotype DX™

The TAILORx trial was designed to establish whether Oncotype DX RS has clinical utility for selection of which patients with node negative hormone receptor-positive breast cancer benefit from receiving chemotherapy in addition to endocrine therapy (predictive ability).

TAILORx: Validation of predictive ability of Oncotype DX RS

9719 eligible patients with follow-up information, 6711 (69%) had a midrange recurrence score of 11 to 25 and were randomly assigned to receive either chemoendocrine therapy or endocrine therapy alone.

Figure from Zujewski & Kamin, Future Oncol 2008;4:603-610



RS Group, therapy	5-yr Invas. DFS (%)	9-yr Invas. DFS (%)
RS≤ 10, Endocrine	94.0 ± 0.6	84.0 ± 1.3
RS 11-25, Endocrine	92.8 ± 0.5	83.3 ± 0.9
RS 11-25, Chemoendocrine	93.1 ± 0.5	84.3 ± 0.8
RS ≥ 26, Chemoendocrine	87.6 ± 1.0	75.7 ± 2.2

Sparano et al, *N Engl J Med* 2015;373:2005-2014
Sparano et al, N Engl J Med 2018;379:111-21.

Summary remarks

- We now have several examples of successful omics-based tests.
- Still a need for education about best practices.
- Successful development and clinical translation takes the right expertise, time, and resources.
- For discussion sessions: How can we translate omics to clinically usefully tests more effectively (highest positive impact for patients) and more efficiently?

THANK YOU!

www.cancer.gov **www.cancer.gov/espanol**