Statistical Analysis of Biomarker Data

Gary M. Clark, Ph.D.

Vice President
Biostatistics & Data Management
Array BioPharma Inc.
Boulder, CO

NCIC Clinical Trials Group New Investigator Clinical Trials Course
August 21-23, 2013

Presenter’s Conflict of Interest

- Gary Clark is a full-time employee of Array BioPharma Inc
- Gary Clark owns stock and has stock options in Array BioPharma Inc
- However, the presentation today reflects the personal opinions of Gary Clark and not necessarily those of Array BioPharma Inc

Clinical Uses for Biomarkers

- Diagnosis/risk assessment

  PSA, BRCA1

- Prognosis/natural history/staging

  Lymph node status, tumor size

- Predicting response to therapy

  ER, PgR, HER-2

- Monitoring response to therapy

  CA125, CA15.3, CEA

- Targets for therapy

  HER-2, EGFR, EML4-ALK fusion gene

Clinical Utility of Biomarkers

“There are few tumor markers that are clinically useful in predicting therapeutic response or patient outcomes despite nearly 20 years of advances in molecular biology.”

Hammond ME, Taube SE, Semin Oncol 2002; 29:213-21

“Even with our ability to identify large numbers of proteins in biofluids for finding clinically useful biomarkers, the discovery and translation of biomarkers for clinical use has been a greater challenge than many expected.”

Waybright TJ, Veenstra TD, Expert Rev Mol Diagn 2009; 9:305-7

Reasons for Conflicting Results in Biomarker Studies

- Different assay protocols or measurement techniques
- Different types of specimens (eg, fresh-frozen vs. fixed tissue, serum)
- Different clinical endpoints (eg, response, DFS, OS)
- Different patient populations (eg, stage, treatments)
- Single study without independent confirmation
- Statistical issues (*eg, Simon R et al, J Natl Cancer Inst 2003; 95:14-8; Lusa L et al, Stat Med 2007; 26:1102-13*)

Biomarker validation begins with validation of the methods (usually, an assay) used to measure the biomarker

Diagnostic Tests

Assay method validation (analytic validation)

- The process of assessing the assay and its measurement performance characteristics
- Determining the range of conditions under which the assay will give reproducible and accurate data

Assay qualification

- The evidentiary process of linking a biomarker with biological processes and clinical endpoints to show that it is “fit for purpose”
- It is dependent on the intended application and it interacts with method validation

Wagner JA et al, Clin Pharmacol Ther 2007; 81:104-7

Assay Method Validation

- Sensitivity and specificity
- Receiver-operating characteristic (ROC) curves
- Positive and negative predictive value
- Positive and negative likelihood ratios
- Overall percent agreement
- Cohen’s kappa

FDA Guidance Document 2007
Guidance for industry and FDA staff: Statistical guidance on reporting results from studies evaluating diagnostic tests
http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm

REporting recommendations for tumor MARKer prognostic studies (REMARK)

McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM for the Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics

- Nat Clin Pract Oncol 2005; 2:416-22
- Eur J Cancer 2005; 41:1690-6
- J Natl Cancer Inst 2005; 97:1180-4
- Br J Cancer 2005; 93:387-91
- J Clin Oncol 2005; 23:9067-72
- Breast Cancer Res Treatment 2006; 100:229-35
- Exp Oncol 2006; 28:99-105

REporting recommendations for tumor MARKer prognostic studies (REMARK): Explanation and Elaboration

Altman DG, McShane LM, Sauerbrei W, Taube SE

- PLoS Med 2012; 9(5) e1001216

  doi: 10.1371/journal.pmed.1001216

- BMC Med 2012 May 29; 10.51

  doi: 10.1186/1741-7015-10-51

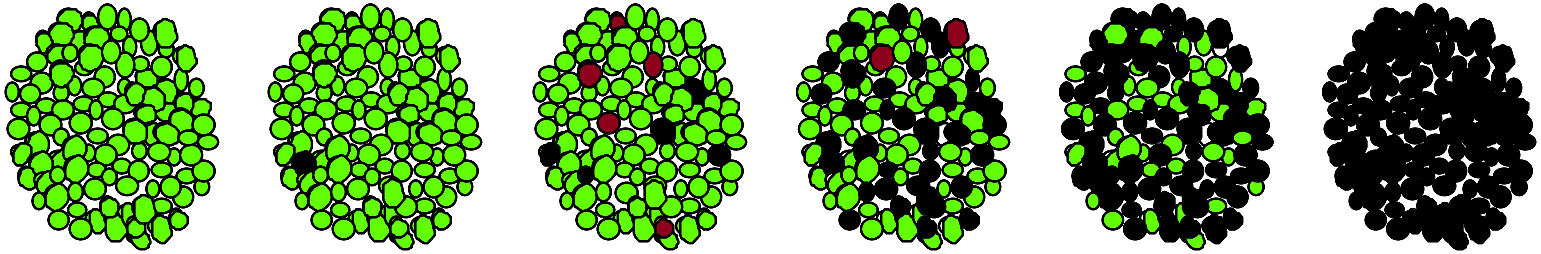
REMARK Guidelines
Materials & Methods

Assay methods

- Specify the assay method used and provide a detailed protocol, including specific reagents or kits used, QC procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols.
- Specify whether and how assays were performed blinded to the study endpoint.

IHC Scoring Systems

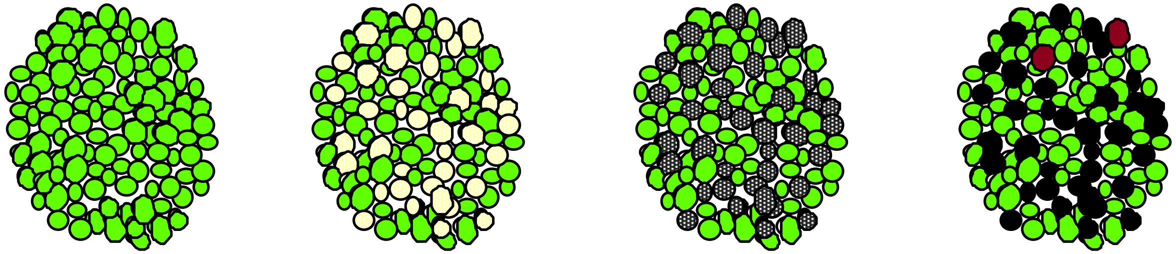
IHC Scoring Systems



0 → 1/100 → 1/10 → 1/3 → 2/3 → 1

Proportion Score (PS) 0 1 2 3 4 5

0 —— 100



Intensity Score (IS) **0** = negative **1** = weak **2** = intermed **3** = strong

Allred DC et al. Mod Pathol 1998; 11:155-68

IHC Scoring Systems

- Percent positive staining (PS) only
- Intensity score (IS) only
- Hybrid scores that combine PS and IS
  - Allred Total Score = PS [0-5] + IS [0-3]
    - Range 0, 2-8
  - Franklin-Hirsch H-Score = PS [0-100] x (IS +1)
    - Range 0 - 400

Effect of Different Definitions of EGFR+ using Dako EGFR pharmDX™ Kits

N = 325 NSCLC Samples[1]

Definition of EGFR+	% EGFR+
Any staining[2]	71%
≥ 10% staining[3]	57%
2+ or 3+ intensity score[4]	47%
H-Score > 200[5]	22%
H-Score > 300[6]	11%

1 Clark et al, J Thorac Oncol 2006; 1:837-46
2 Dako EGFR pharmDX™ kit
3 Tsao MS et al, N Engl J Med 2005; 353:133-44
4 Pérez-Soler R et al, J Clin Oncol 2004; 22:3238-47
5 Hirsch FR et al, J Clin Oncol 2003; 21:3798-804
6 Cappuzzo F et al, J Natl Cancer Inst 2005; 97:643-55

REMARK Guidelines
Materials & Methods

Statistical analysis methods

- Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.
- Clarify how marker values were handled in the analyses.
- If relevant, describe methods used for cutpoint determination.

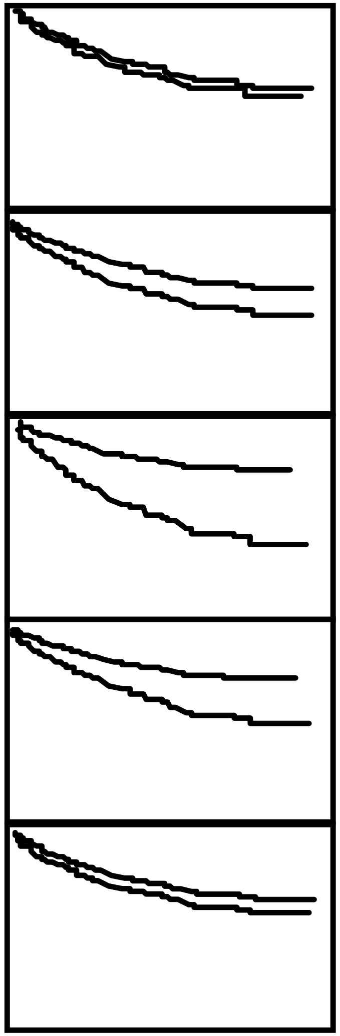
CUTPOINT ANALYSES

How to Select a Cutpoint

Cutpoints that frequently appear in the literature

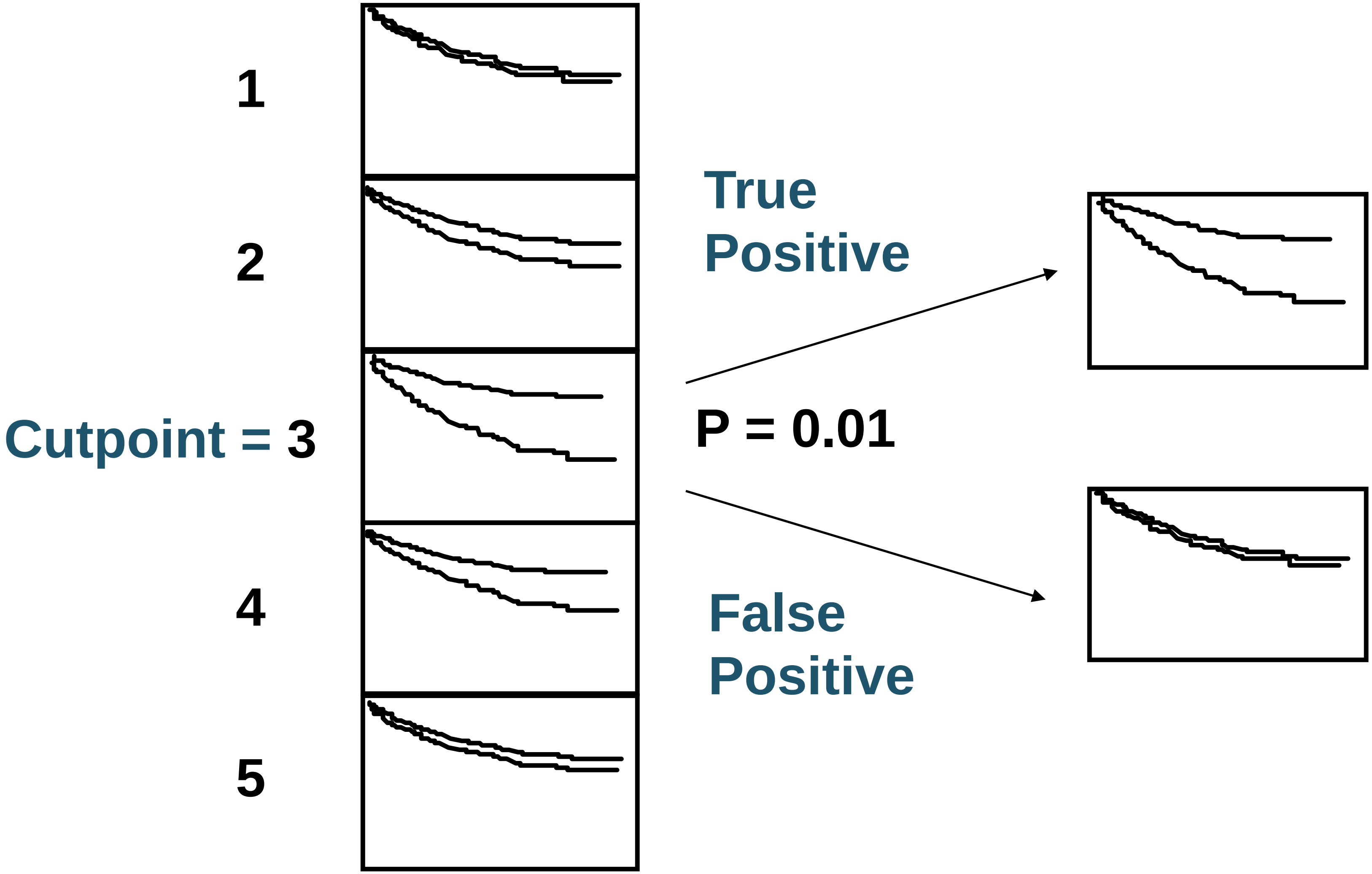
- Median
- Lower or upper quartile
- A value from the literature
- An “optimal” cutpoint based on correlation with clinical outcome

Find the “Best” Cutpoint



Cutpoint		
1	P = 0.70	
2	P = 0.30	
3	P = 0.01	← Best cutpoint
4	P = 0.08	
5	P = 0.50	

Validation of the "Best" Cutpoint



1

2

Cutpoint = 3

4

5

True Positive

P = 0.01

False Positive

Simulation Study

Investigate the problem of false positives when continuous biomarkers are dichotomized

Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? Breast Cancer Res Treat 1992; 22:197-206

Simulation Experimental Design

- 250 simulated patients in each dataset
- Randomly generated biomarker values
- Randomly generated recurrence times

5 yr DFS = 70%

Average follow-up = 6 years

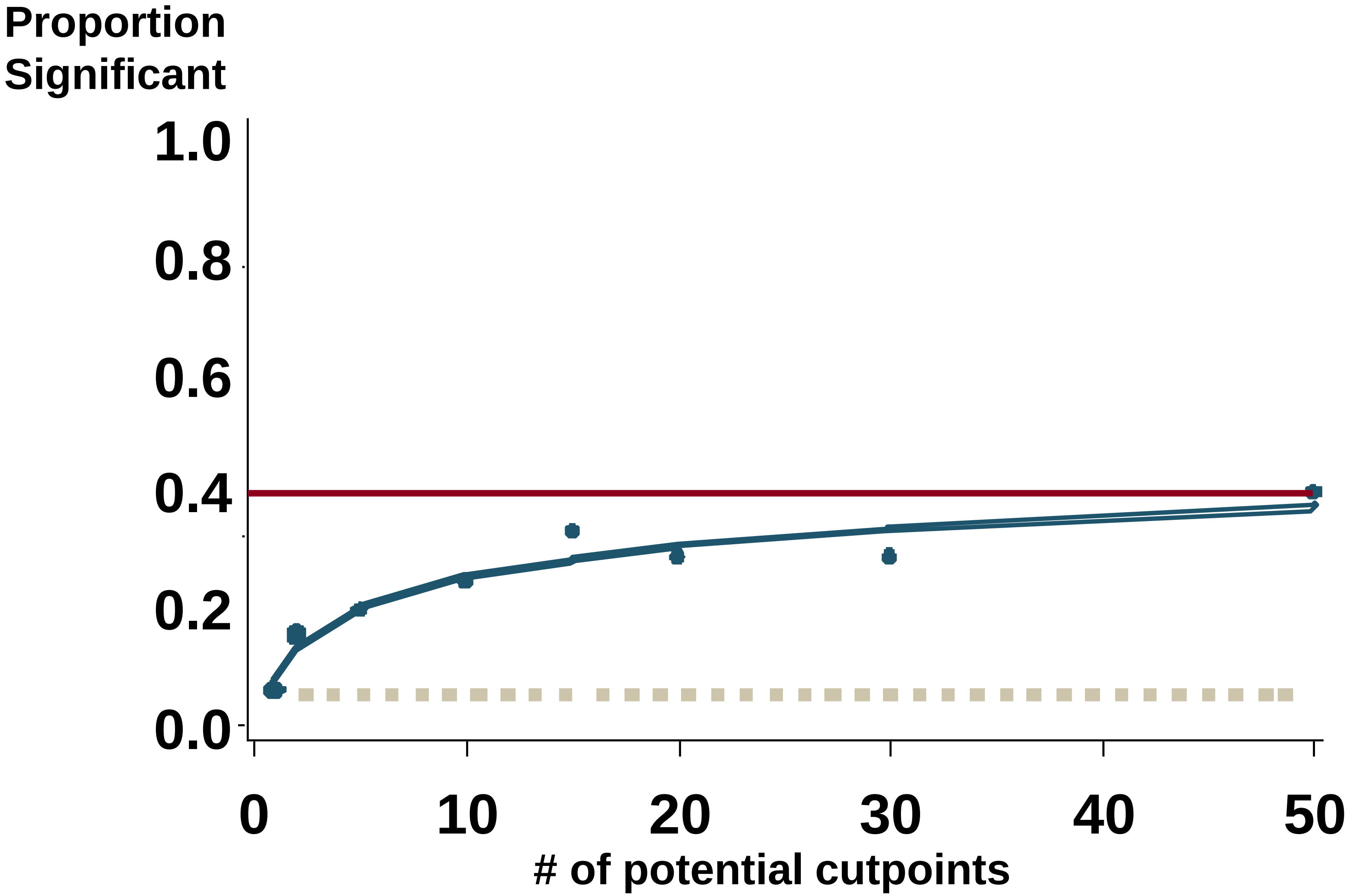
Experimental Design

Number of possible values

		1	2	5	10	15	20	30	50	
Actual Improvement = in DFS	0%	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	Null

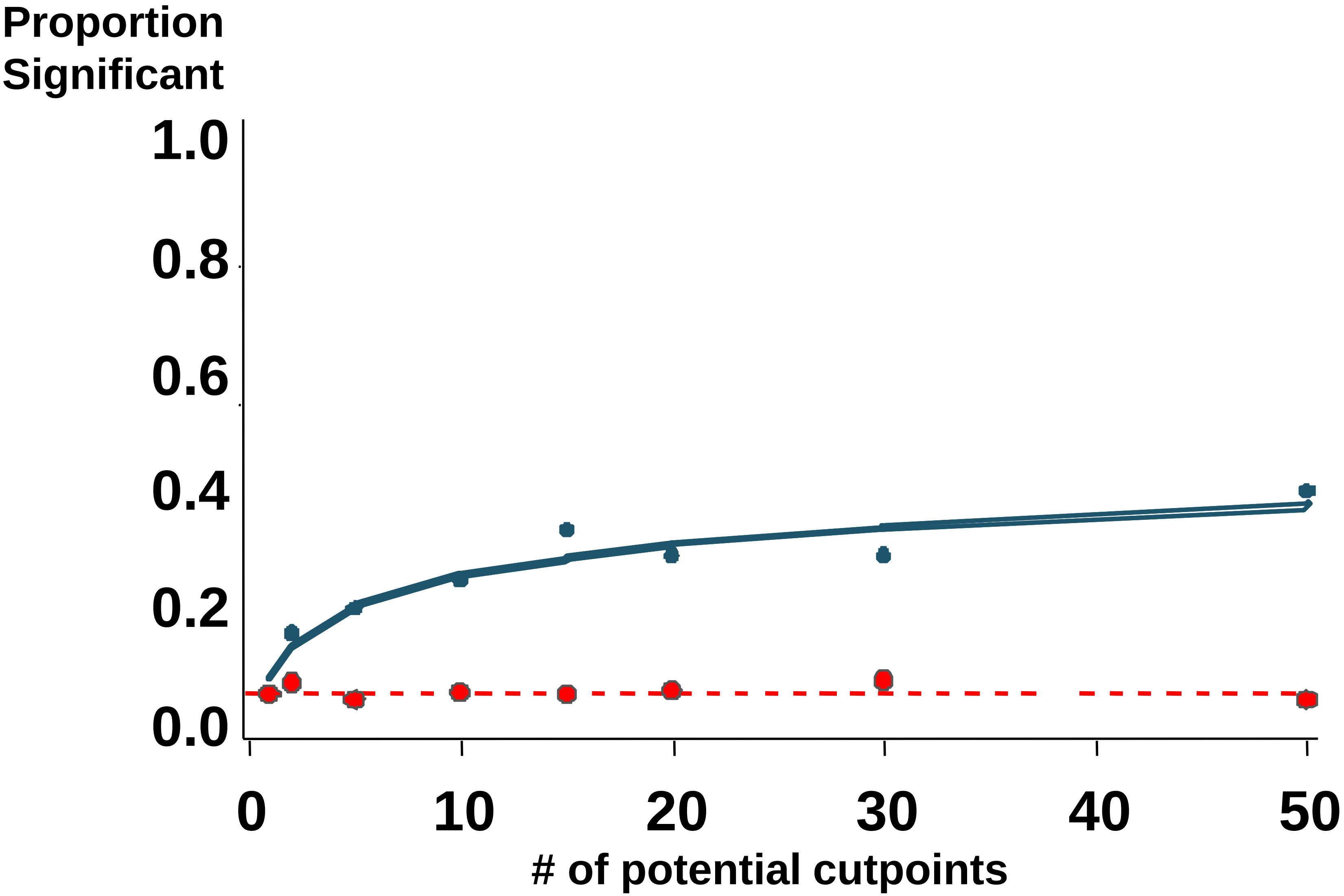
- Run each scenario 200-300 times
- Calculate log rank P value for Marker+ vs. Marker-
- Count number of runs with a cutpoint $P < 0.05$

Null Hypothesis of NO Effect is True



Null Hypothesis of NO Effect is True

Validation Results



Approaches to the Problem

- Separate training and validation sets of data
- Internal validation
  - Jack knife
  - Bootstrap
- P-value adjustment

Conclusions

Even modest data exploration runs a serious risk of:

- Finding an effect where none exists
- Overestimating the importance of a new biomarker

Validation is essential!

What about multiplex assays or gene expression profiles?

- Many, many genes
- Big problem of multiple hypothesis testing
- Gene expression profiling tests are considered medical devices by the FDA

Gene Expression Profile Tests for Early Stage Breast Cancer

Effectiveness Guidance Document:

Methodological Guidance for the Design of Comparative Effectiveness Studies

Center for Medical Technology Policy

Version 1.0 Published June 2009

www.cmtpnet.org/effectiveness-guidance-documents/gene-expression-predictors-for-breast-cancer-egd

Gene Expression Profile Tests for Early Stage Breast Cancer

Examples of recommendations

- The manner in which the test or algorithm is developed has little or no relevance to approval standards, as the latter are based almost exclusively on external validation ("test" set) results
- The population used to validate the prognostic algorithm (the "test" set) must be completely independent from the one used to develop the algorithm (the "training" set)
- The test, including the complete algorithm, created in the development or discovery phase cannot be altered in the validation phase (if it is, additional independent data must be used to validate the modified algorithm)
- Some of the validation must be of the entire test procedure (ie, not just the performance of the expression "signature" as measured in research settings); this means that patient samples must be sent, as they would in clinical practice, to the same lab and subject to the same procedures as will be used for the marketed test

In Real Estate:

- Location, Location, Location!

In Biomarker Research:

- Validation, Validation, Validation!