Statistical Methods and Statistical Pitfalls in Biomarker Research

Frank E Harrell Jr

Department of Biostatistics, Vanderbilt University School of Medicine

VU Biomarker Research Summit 22 June 2007

Outline

Bad Epidemiologic Practice

Biases might pose a special challenge for laboratory researchers who are used to biological reasoning and the tightly controlled conditions of experimental research. Such researchers unwittingly become non-experimental observational epidemiologists when they apply molecular assays in studies of diagnosis and prognosis, for which the experimental method is not available and for which biological reasoning might have limited usefulness.

Ransohoff [2005]; see also Ransohoff [2004]

Bad Statistical Practice

- Data torture
- Subsetting subjects
- Finding genes using subjects to later be used in independent validation
- Analyzing time-to-event data as binary responses
- Choosing cutpoints to optimize accuracy
- Incorrect accuracy measures
- Incomplete or no validation
- Overstatement of results

Bad Statistical Practice, *cont.*

- No demonstration that information is new; not giving clinical variables same opportunities as potential biomarkers
- Poor use of continuous markers
- Failure to use fully reproducible scripted data management and analysis
- Presenting only the result that validates best
- See REMARK guidelines [McShane et al., 2005], Ioannidis [2007], Biostatistics Web [2007]

Statistical Goals

- Experimental design, e.g. randomize processing order, blinding to patient outcome
- Understanding the measurements
- Analyzing assay variability/reliability
- Normalization (**better**: build into comprehensive model)
- Finding diagnostically or prognostically useful biomarkers
- Determines appropriate transformations
- Demonstrating reproducible signal
- Unbiased validation of predictive accuracy
- Demonstrating information added to cheap clinical variables
- Interpretation: risk plots, nomograms

Demonstration of Added Information

- Biomarkers must add information to already available information
  - Partial test of association controlling for cheap info
  - Index of information gain
- Show that biomarker values cannot be predicted from existing data
- Insufficient number of cases to adjust for many clinical variables $\longrightarrow$ propensity score analysis
  - Predict marker value from all clinical variables
  - Solely adjust for predicted marker value

Difficulties of Picking "Winners"

- Multiple comparison problems
- Extremely low power; high false negative rate
- Potential markers may be correlated with each other
- Small changes in the data can change the winner
- Significance testing can be irrelevant; is a ranking and selection problem

Ranking Markers

- Bootstrap (Efron): simulate performance of a statistic by resampling (with replacement) from your data
- Can use it to solve difficult problems, e.g. confidence interval for the number of modes in a distribution
- Useful here for quantifying information in the dataset for picking winners
  - Attempt to rank competing markers by a test statistic (crude or **partial**)
  - Compute 0.95 confidence intervals of ranks—stability of observed rank

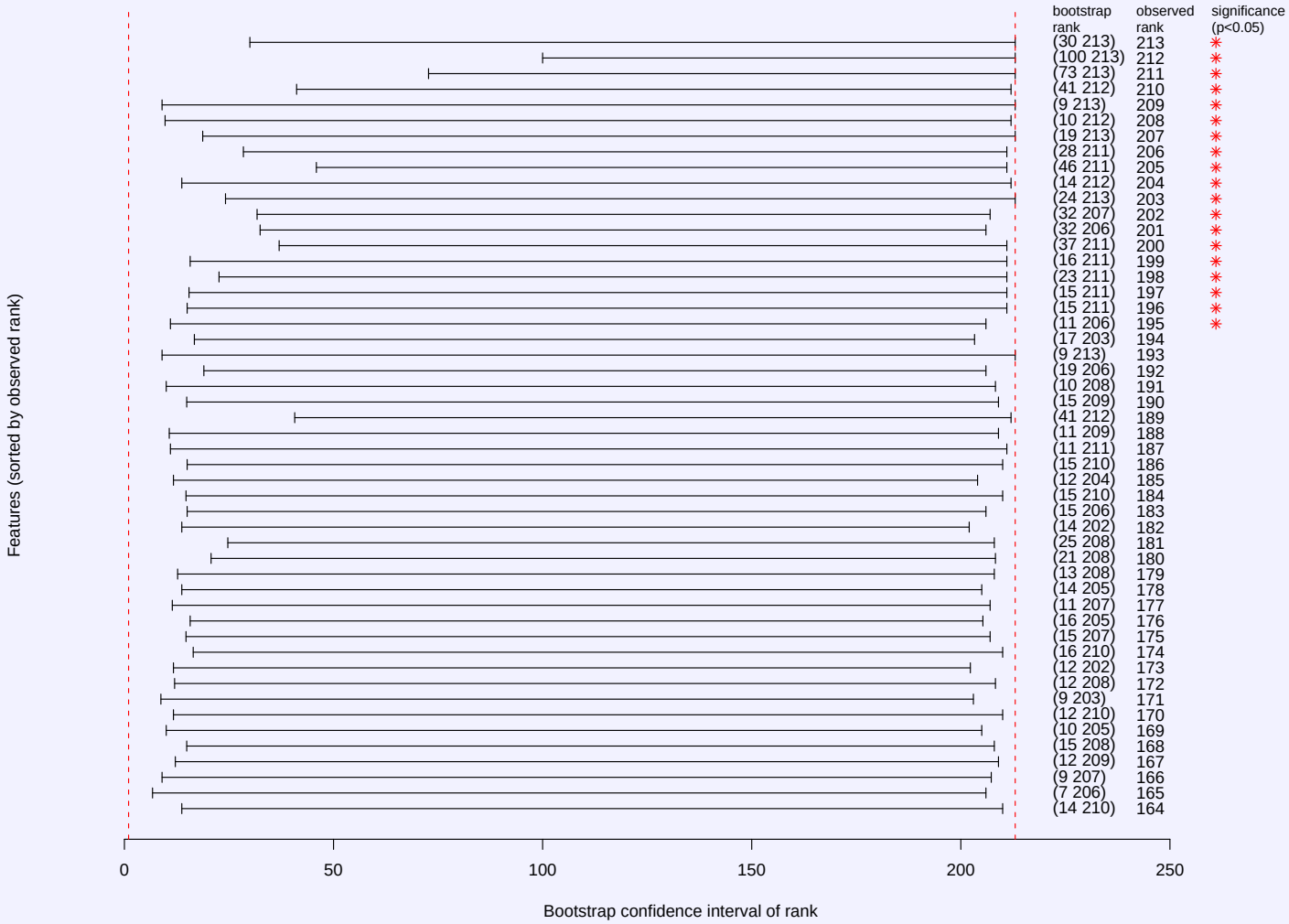
Example

- Research led by Michael Edgeworth (Neurology) and Richard Caprioli
- Analysis done by Deming Mi M.S. Dept. of Biostatistics and Mass Spec Research Lab
- Tissue samples from 54 patients, 0.63 of them died
- Malignant glioma, receiving post-op chemotherapy
- Cox model adjusted for age, tumor grade, radiation
- Median follow-up 15.5m for survivors
- Median survival 15m

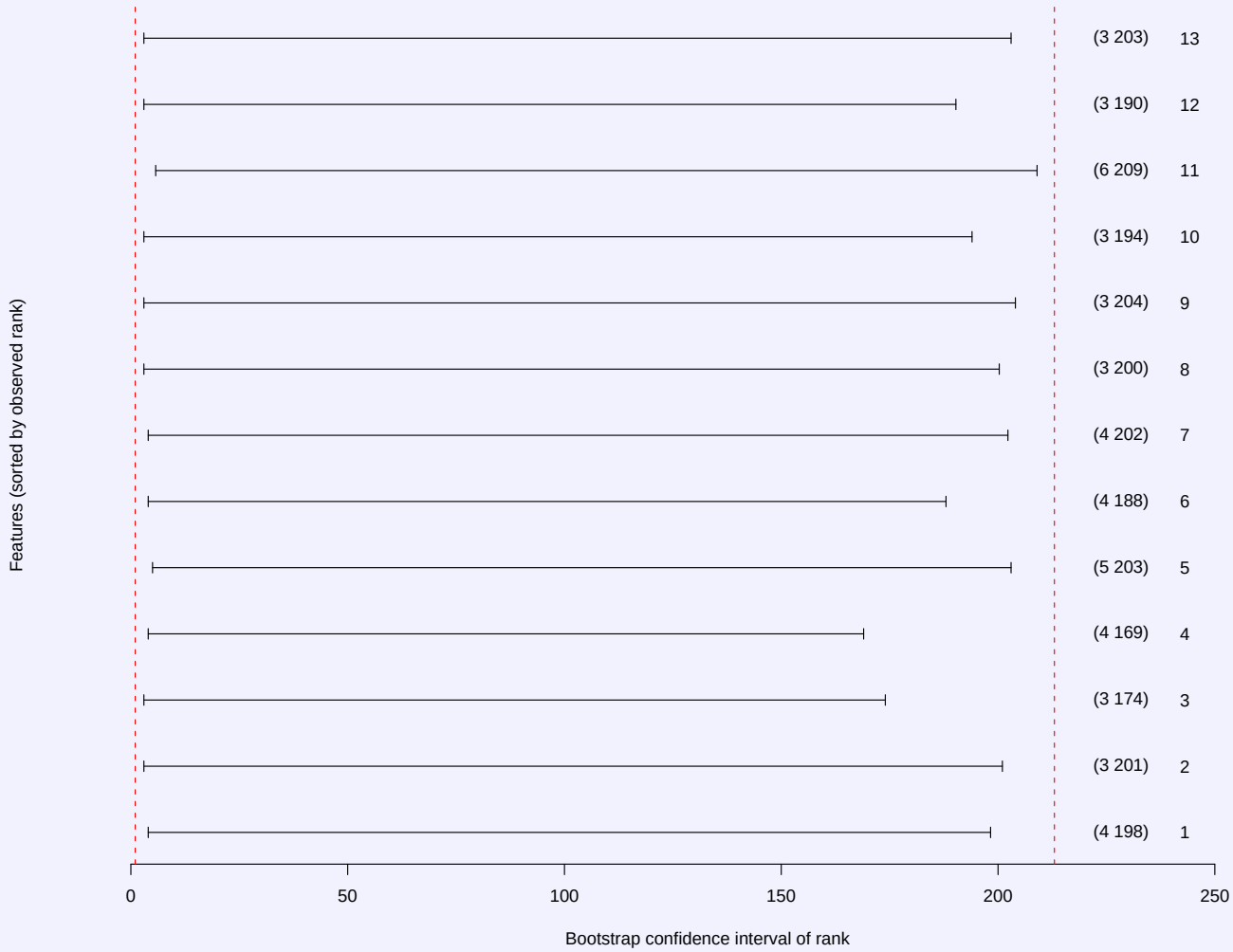
Example, *cont.*

- 213 candidate features extracted from avg. spectrum using ProTS-Marker (Biodesix Inc.)
- Ranked by partial likelihood ratio χ^2
- 600 resamples from original data, markers re-ranked each time
- 0.025 and 0.975 quantiles of ranks
- Features sorted by observed ranks in the whole sample
- Significant associations have asterisks

Results - Best



Results - Worst



Problems with Classification

- Proportion classified correctly is an **improper scoring rule**
  - Optimized by bogus model
- Minimum information
  - low statistical power
  - high standard errors of regression coefficients
  - arbitrary to choice of cutoff on predicted risk
  - forces binary decision, does not yield a "gray zone" $\rightarrow$ more data needed
- Assumes statistician to be provider of utility function
- Sensitivity and specificity are also improper scoring rules

Example: Damage Caused by Improper Scoring Rule

- Predicting probability of an event, e.g., Prob(disease)
- $N = 400$, 0.57 of subjects have disease
- Classify as diseased if prob. > 0.5

Model	C Index	χ^2	Proportion Correct
age	.592	10.5	.622
sex	.589	12.4	.588
age+sex	.639	22.8	.600
constant	.500	0.0	.573

Adjusted Odds Ratios:

age (IQR 58y:42y)	1.6 (0.95CL 1.2-2.0)
sex (f:m)	0.5 (0.95CL 0.3-0.7)

Need for Stringent Validation

- Splitting a sample does not provide external validation
- Split-sample validation is terribly inefficient and arbitrary unless $>$ 20,000 subjects
- Greater reliability obtained by using all subjects and using bootstrap or 50 repeats of 10-fold cross validation
- Must repeat **ALL** steps that were unblinded to outcome variable for each re-sample
- Use a proper scoring rule (e.g., Brier score, logarithmic score) or correlation between predicted risk and observed outcome (R^2 or rank correlation–concordance index such as ROC area)
- ROC area is not good for comparing two models [Pencina et al., 2007, Peek et al., 2007]
- Necessary to unbiasedly validate a high-resolution calibration curve (smooth plot of predicted vs. actual risk of outcome)

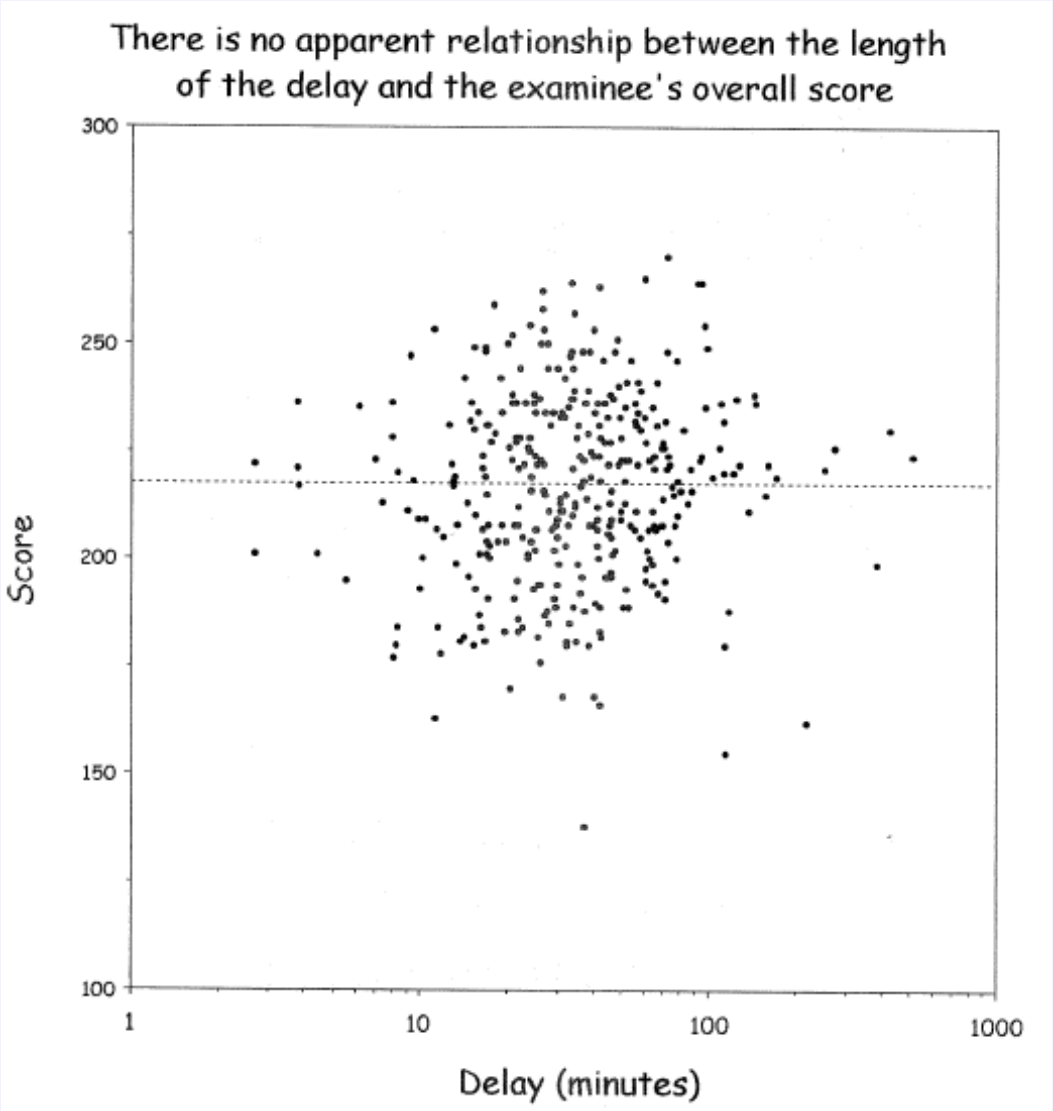
Problems Caused by Chopping Continuous Variables

- Chopping predicted probabilities causes major problems
- Many problems caused by chopping predictors
- True cutpoints do not exist unless risk relationship discontinuous
- Cutpoints may be found that result in both increasing and decreasing relationships with **any** dataset with zero correlation

Range of Delay	Mean Score	Range of Delay	Mean Score
0-11	210	0-3.8	220
11-20	215	3.8-8	219
21-30	217	8-113	217
31-40	218	113-170	215
41-	220	170-	210

Wainer [2006]; See "Dichotomania" [Senn, 2005] and Royston et al. [2006]

Data from Wainer [2006]



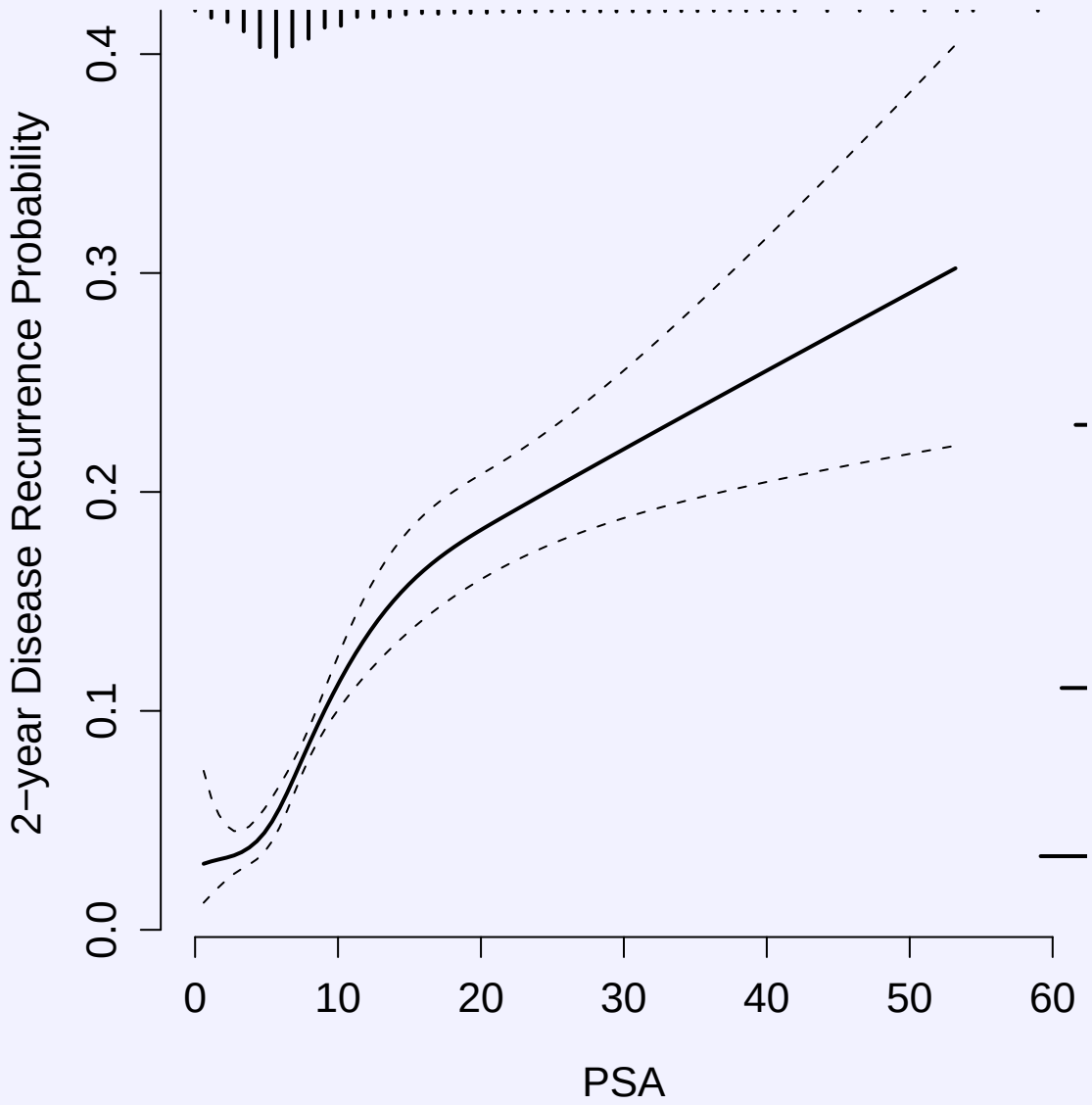
Cutpoints are Disasters

. . . in almost every study where [finding optimal cutpoints] is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the 'optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were in included with 12 different cutpoints . . . Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of ASCO. —*Holländer et al. [2004]*

Value of Continuous Biomarkers

- Avoid arbitrary cutpoints
- Better risk spectrum
- Provides gray zone
- Increases power/precision
- Fewer biomarkers required to achieve same accuracy
- $\longrightarrow$ prediction rules are simpler

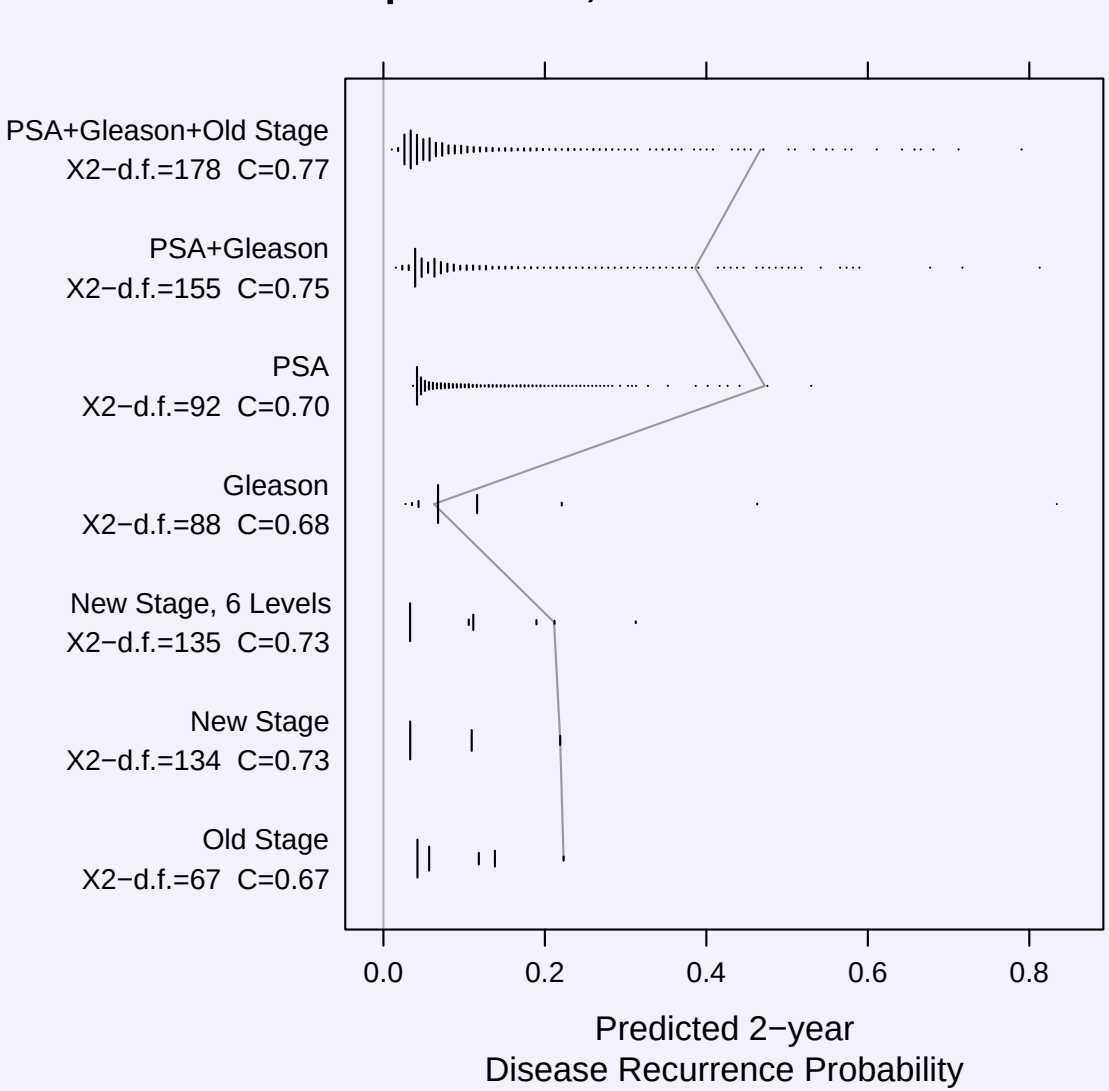
Prognosis in Prostate Cancer



Data courtesy of M Kattan from JNCI 98:715; 2006

Horizontal ticks represent frequencies of prognoses by new staging system

Prognosis in Prostate Cancer, *cont.*



Prognosis after Myocardial Infarction

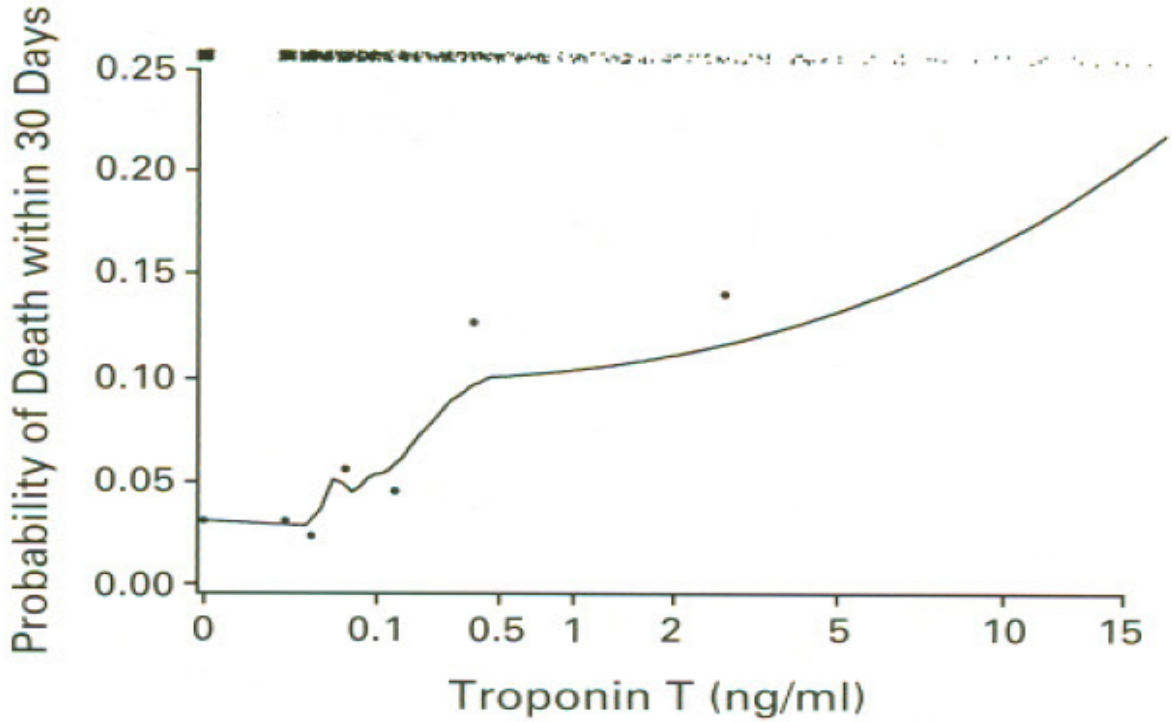


Figure 2. Probability of Death within 30 Days According to the Troponin T Level at Hospital Admission.

Smoothed nonparametric estimates are shown. The troponin T levels are plotted on a cube-root scale. The density of the data is indicated at the top, with each mark representing one patient. The dots represent simple estimates of mortality derived from ranges of the troponin T level that contained at least 70 patients.

Ohman et al. [1996]

Summary

- Current state of biomarker analysis leaves much to be desired
- Many statistical and epidemiologic problems, especially:
  - bias
  - overfitting and overstatement
  - incomplete validation
  - loss of information and $\uparrow$ arbitrariness caused by chopping continuous quantities
  - misleading results based on classification accuracy
  - failure to adjust for cheap information
- Cutpoints are inherently misleading
- Picking winners $\equiv$ splitting hairs
- Analyze clinical data as aggressively as potential biomarkers

References

Biostatistics Web. Checklist for authors, 2007. `http://biostat.mc.vanderbilt.edu/ManuscriptChecklist`.

N. Holländer, W. Sauerbrei, and M. Schumacher. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Stat Med*, 23:1701–1713, 2004.

J. P. A. Ioannidis. Is molecular profiling ready for use in clinical decision making? *The Oncologist*, 12: 301–311, 2007.

L. M. McShane, D. G. Altman, W. Sauerbrei, S. E. Taube, M. Gion, G. M. Clark, and The Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Nat Cancer Inst*, 97:1180–1184, 2005.

E. M. Ohman, P. W. Armstrong, R. H. Christenson, C. B. Granger, H. A. Katus, C. W. Hamm, M. A. O'Hannesian, G. S. Wagner, N. S. Kleiman, F. E. Harrell, R. M. Califf, E. J. Topol, K. L. Lee, and the GUSTO-IIa Investigators. Cardiac troponin T levels for risk stratification in acute myocardial ischemia. *NEJM*, 335:1333–1341, 1996.

N. Peek, D. G. T. Arts, R. J. Bosman, P. H. J. van der Voort, and N. F. de Keizer. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epi*, 60:491–501, 2007.

M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, and R. S. Vasan. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med*, 26, 2007.

D. F. Ransohoff. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev*, 4: 309–314, 2004.

D. F. Ransohoff. Bias as a threat to validity of cancer molecular-marker research. *Nat Rev*, 5:142–149, 2005.

P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25:127–141, 2006.

S. J. Senn. Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. In *Proceedings of the International Statistical Institute, 55th Session*, Sydney, 2005.

H. Wainer. Finding what is not there through the unfortunate binning of results: The Mendel effect. *Chance*, 19(1):49–56, 2006.