How To Do Bad Biomarker Research

Frank E Harrell Jr

Department of Biostatistics, Vanderbilt University School of Medicine

Towards Building Better Biomarkers
Statistical Methodology

NIDDK NIH Bethesda MD 2014-12-02

Vanderbilt Center for Quantitative Sciences Workshop 2015-01-16
Fundamental Principles of Statistics

- Use methods grounded in theory or extensive simulation
- Understand uncertainty
- Design experiments to maximize information
- Use all information in data during analysis
- Use discovery and estimation procedures not likely to claim that noise is signal
- Give decision makers the inputs (other than the utility function) that optimize decisions
To Treat Depression, Drugs or Therapy?

By RICHARD A. FRIEDMAN, M.D.  JANUARY 6, 2015 8:00 AM  236 Comments

You’re feeling down, and your doctor or therapist has confirmed it: You have depression. Now what?

Until recently, many experts thought that your clinician could literally pick any antidepressant or type of psychotherapy at random because, with a few clinical exceptions, there was little...
Original Investigation

Toward a Neuroimaging Treatment Selection Biomarker for Major Depressive Disorder

Callie L. McGrath, BA; Mary E. Kelley, PhD; Paul E. Holtzheimer III, MD; Boadie W. Dunlop, MD; W. Edward Craighead, PhD; Alexandre R. Franco, PhD; R. Cameron Craddock, PhD; Helen S. Mayberg, MD

**IMPORTANCE** Currently, fewer than 40% of patients treated for major depressive disorder achieve remission with initial treatment. Identification of a biological marker that might improve these odds could have significant health and economic impact.

**OBJECTIVE** To identify a candidate neuroimaging “treatment-specific biomarker” that predicts differential outcome to either medication or psychotherapy.

Investigators claimed to learn from complex brain imaging analysis on 6 non-responders to drug and 9 non-responders to behavioral therapy.
<table>
<thead>
<tr>
<th>Variable</th>
<th>CBT</th>
<th>Escitalopram Oxalate</th>
<th>Significance of Group × Treatment Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, y, mean (SD)</td>
<td>45.4 (8.8)</td>
<td>42.5 (10.8)</td>
<td>40.3 (5.2)</td>
</tr>
<tr>
<td>Male, No. (%)</td>
<td>3 (33.3)</td>
<td>7 (58.3)</td>
<td>2 (33.3)</td>
</tr>
<tr>
<td>White, No. (%)</td>
<td>7 (77.8)</td>
<td>8 (66.7)</td>
<td>5 (100)</td>
</tr>
<tr>
<td>Education, y, mean (SD)</td>
<td>14.7 (1.7)</td>
<td>15.6 (1.6)</td>
<td>15.4 (1.7)</td>
</tr>
<tr>
<td>Age at onset of MDD, y, mean (SD)</td>
<td>28.4 (12.3)</td>
<td>28.7 (11.2)</td>
<td>24.0 (11.6)</td>
</tr>
<tr>
<td>≥3 Lifetime episodes, No. (%)</td>
<td>4 (50.0)</td>
<td>4 (33.3)</td>
<td>2 (33.3)</td>
</tr>
<tr>
<td>Duration of current episode, wk, mean (SD)</td>
<td>124.5 (118.5)</td>
<td>257.3 (308.8)</td>
<td>299.8 (620.0)</td>
</tr>
<tr>
<td>No. of previous AD trials in current episode, mean (SD)</td>
<td>1.2 (1.5)</td>
<td>1.2 (1.0)</td>
<td>1.5 (1.4)</td>
</tr>
<tr>
<td>Melancholic subtype, No. (%)</td>
<td>4 (44.4)</td>
<td>5 (45.5)</td>
<td>3 (50.0)</td>
</tr>
<tr>
<td>Current anxiety disorder, No. (%)</td>
<td>3 (33.3)</td>
<td>2 (16.7)</td>
<td>1 (16.7)</td>
</tr>
<tr>
<td>Baseline HDRS score, mean (SD)</td>
<td>19.9 (3.8)</td>
<td>17.9 (2.7)</td>
<td>18.0 (2.1)</td>
</tr>
<tr>
<td>Baseline HAMA total score, mean (SD)</td>
<td>18.9 (7.6)</td>
<td>12.8 (2.8)</td>
<td>13.3 (2.3)</td>
</tr>
<tr>
<td>BDI total score, mean (SD)</td>
<td>19.2 (4.6)</td>
<td>18.7 (7.2)</td>
<td>20.0 (3.8)</td>
</tr>
</tbody>
</table>

McGrath CL et al, JAMA Psychiatry 2013
Expanded view of findings. A, Scatterplot of insular activity from individual subjects in the remitter (REM) and nonresponder (NR) groups. Note: the anterior insula is the only region where the interaction subdivides patients into hypermetabolic (region/whole-brain mean >1.0) and hypometabolic (region/whole-brain mean <1.0) subgroups. B, Correlations of insula activity with percentage of change in Hamilton Depression Rating Scale (HDRS) score in the full cohort of subjects treated with cognitive behavior therapy (CBT) and escitalopram oxalate.
The Palin Effect: Avoiding Information

If you like any of these physician responses we can stop here . . .

**Patient:** What was my systolic BP this time?

**MD:** It was > 120

**Patient:** How is my diabetes doing?

**MD:** Your Hb\(_{A1c}\) was > 6.5

**Patient:** What about the prostate screen?

**MD:** If you have average prostate cancer, the chance that PSA > 5 in this report is 0.6

**Problem:** Improper conditioning (\(X > c\) instead of \(X = x\)) \(\rightarrow\) information loss; reversing time flow

**Sensitivity:** \(P(\text{observed } X > c \text{ given unobserved } Y = y)\)
Goals

Create a diagnostic or prognostic model that
- will be of limited clinical utility
- will not be strongly validated in the future
- has an interpretation that is not what it seems
- uses cut-points, when cut-points don’t even exist, that
  - others will disagree with
  - result in discontinuous predictions and thinking
  - requires more biomarkers to make up for information loss due to dichotomization

Find a biomarker to personalize treatment selection that is not as reliable as using published average treatment effects from RCTs
Study Design

- Ignore the clinical literature when deciding upon clinical variables to collect
  - Don’t allow clinical variables to have dimensionality as high as the candidate biomarkers
- Don’t randomize the order of sample processing; inform lab personnel of patient’s outcome status
- Don’t study reliability of biomarker assays or clinical variables
- Re-label demographic variables as clinical variables
- Choose a non-representative sample
- Double the needed sample size by dichotomizing the outcome measure
- Reduce the effective sample size by splitting into training and validation samples
Statistical Analysis Plan

- Don’t have one as this might limit investigator flexibility to substitute hypotheses
- Categorize continuous variables or assume they operate linearly
- Even though the patient response is a validated continuous measurement, analyze it as “high” vs. “low”
- Use univariable screening and stepwise regression
- Ignore time in time-to-event data
- Choose a discontinuous improper predictive accuracy score to gauge diagnostic or prognostic ability
- Try different cut-points on all variables in order to get a good value on the improper accuracy score
- Use Excel or a menu-based statistical package so no one can re-trace your steps in order to criticize them
Interpretation and Validation

- Pretend that the clinical variables you adjusted for were adequate and claim that the biomarkers provide new information
- Pick a “winning” biomarker even though tiny changes in the sample result in a different “winner”
- Overstate the predictive utility in general
- Validate predictions using an independent sample that is too small or that should have been incorporated into training data to achieve adequate sample size
- If the validation is poor, re-start with a different data split and proceed until validation is good
- Avoid checking the absolute accuracy of predictions; instead group predictions into quartiles and show the quartiles have different patient outcomes
- Categorize predictions to avoid making optimum Bayes decisions
Vanderbilt-led team studies blood test for prostate cancer

by Bill Snyder | Posted on Monday, Jan. 5, 2015 — 9:39 AM

Vanderbilt University researcher William Mitchell, M.D., Ph.D., and colleagues in Germany and Canada have demonstrated a method for detecting “cell-free” tumor DNA in the bloodstream.

Mitchell believes the technique will be transformative in providing improved cancer diagnostics that can both predict treatment outcomes and monitor patient responses to therapy.

In a large retrospective study of blood samples, the researchers showed that the method, called a “liquid biopsy,” could accurately distinguish prostate cancer from normal controls without prior knowledge of the genetic “signature” of the tumors, and with over three times the sensitivity of current prostate-specific antigen (PSA) screening.

The study appears in the January issue of *Clinical Chemistry* (volume 61, page 239), which is dedicated to “Molecular Diagnostics: A Revolution in Progress.”

“Based on the reported data and work in progress, I believe the ‘liquid biopsy’ will revolutionize cancer diagnostics, not only before a patient begins therapy but also following patient responses to therapy,” said Mitchell, the paper’s corresponding author and professor of Pathology, Microbiology and Immunology.

The study collected serum from more than 200 patients with prostate cancer and more than 200 controls. The samples included PSA levels and prostate tissue biopsy grading, called the Gleason score.

The researchers reported that the technique distinguished prostate cancer from normal controls with 84-percent accuracy and cancer from benign hyperplasia and prostatitis with an accuracy of 91 percent.
Damage Caused by Improper Scoring Rule

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

<table>
<thead>
<tr>
<th>Model</th>
<th>( c ) Index</th>
<th>( \chi^2 )</th>
<th>Proportion Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>.592</td>
<td>10.5</td>
<td>.622</td>
</tr>
<tr>
<td>sex</td>
<td>.589</td>
<td>12.4</td>
<td>.588</td>
</tr>
<tr>
<td>age+sex</td>
<td>.639</td>
<td>22.8</td>
<td>.600</td>
</tr>
<tr>
<td>constant</td>
<td>.500</td>
<td>0.0</td>
<td>.573</td>
</tr>
</tbody>
</table>

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)
**Thresholds**

**Can Occur in Biology**
Not Handled by Dichotomization

**Cannot Occur Unless** $X = \text{time}$
Assumed in Much of Biomarker Research

![Graphs showing the difference between can occur in biology and cannot occur unless a specific condition is met.](image-url)
“The treatment is called successful if either the patient has gone down from a baseline diastolic blood pressure of \( \geq 95 \text{ mmHg} \) to \( \leq 90 \text{ mmHg} \) or has achieved a 10\% reduction in blood pressure from baseline.”

Is a mean difference of 5.4\text{mmHg} more difficult to interpret than A: 17\% vs. B: 22\% hit clinical target?
Mathematical Impossibility of Fixed Cutpoints

Diagnosis of Pneumonia in Sick Children 42–90 Days Old

Adjusted Respiratory Rate/min.
Probability of Pneumonia

- Cough
- No cough

References
Apparent Thresholds are Artifacts

In + studies: threshold 132–800 ng/L, correlation with study median $r = 0.86$.

Giannoni et al. [2014]
Problems in Defining AKI and CKD

- Definitions based on information-losing categorization of powerful continuous variables
- SCr interacts less strongly with sex and race when predicting mortality $\rightarrow$ eGFR may be less prognostic than SCr
- SCr vs. mortality and diabetes is a hockey stick
- Definitions based on change ignore (1) hockey stick (2) patient’s current state more important than path
- For critically ill patients in ICU who survived past first 3 days, day 3 SCr has 0.92 of the information of day 1 and day 3 SCr combined in predicting death or SCr $> 2.5$ by day 7
Baseline SCr and Prognosis

Estimated risk for 7772 critically ill ICU patients with day 1 SCr < 2
(SUPPORT, Knaus et al. [1995])

\[
\text{Estimated risk for 7772 critically ill ICU patients with day 1 SCr < 2}
\]

(SUPPORT, Knaus et al. [1995])

- **Hospital Death**
  - **Female**
  - **Male**

**Serum Creatinine Day 3, mg/dl**

- **Hospital Death**
  - **Sex**
  - **Male**
  - **Female**

**CKD – EPI eGFR, ml/min/1.73m²**

- **Hospital Death**
  - **Sex**
  - **Male**
  - **Female**

**Glycohemoglobin, %**

- **Male**
- **Female**

**NHANES 2009–2010**

- **Creatinine, mg/dL**
  - **Male**
  - **Female**

**eGFR requires complex interaction with sex**

**Predictive value:** \( c = 0.61 \text{ SCr}, 0.57 \text{ eGFR} \)
Most Important Predictors of SCr in NHANES

- Age, years
  - $\chi^2_8 = 552.4$

- Arm Circumference, cm
  - $\chi^2_4 = 103.3$

- Standing Height, cm
  - $\chi^2_4 = 82.2$

- Waist Circumference, cm
  - $\chi^2_4 = 84$

- Race/Ethnicity
  - Mexican American
  - Non-Hispanic Black
  - Non-Hispanic White
  - Other Hispanic
  - Other Race Including Multi-Racial
    - $\chi^2_4 = 352.7$

- Sex
  - Male
  - Female
    - $\chi^2_5 = 764.8$
What’s Gone Wrong with Omics & Biomarkers?

- Subramanian and Simon [2010]: Gene expression-based prognostic signatures in lung cancer: Ready for clinical use?
- NSCLC gene expression studies 2002–2009, $n \geq 50$
- 16 studies found
- Scored on appropriateness of protocol, stat validation, medical utility
- Average quality score: 3.1 of 7 points
- No study showed prediction improvement over known risk factors; many failed to validate
- Most studies did not even consider factors in guidelines
  - Completeness of resection only considered in 7
  - Similar for tumor size
  - Some only adjusted for age and sex
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 re-sampling (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.
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
213 candidate features extracted from avg. spectrum using ProTS-Marker (Biodesix Inc.)

- Ranked by partial likelihood ratio $\chi^2$
- 600 re-samples 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

Bootstrap confidence interval of rank
Features (sorted by observed rank)

<table>
<thead>
<tr>
<th>bootstrap rank</th>
<th>observed rank</th>
<th>significance (p&lt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(30, 213)</td>
<td>213</td>
<td>*</td>
</tr>
<tr>
<td>(100, 213)</td>
<td>212</td>
<td>*</td>
</tr>
<tr>
<td>(73, 213)</td>
<td>211</td>
<td>*</td>
</tr>
<tr>
<td>(41, 212)</td>
<td>210</td>
<td>*</td>
</tr>
<tr>
<td>(9, 213)</td>
<td>209</td>
<td>*</td>
</tr>
<tr>
<td>(10, 212)</td>
<td>208</td>
<td>*</td>
</tr>
<tr>
<td>(19, 213)</td>
<td>207</td>
<td>*</td>
</tr>
<tr>
<td>(28, 211)</td>
<td>206</td>
<td>*</td>
</tr>
<tr>
<td>(46, 211)</td>
<td>205</td>
<td>*</td>
</tr>
<tr>
<td>(14, 212)</td>
<td>204</td>
<td>*</td>
</tr>
<tr>
<td>(24, 213)</td>
<td>203</td>
<td>*</td>
</tr>
<tr>
<td>(32, 207)</td>
<td>202</td>
<td>*</td>
</tr>
<tr>
<td>(32, 206)</td>
<td>201</td>
<td>*</td>
</tr>
<tr>
<td>(37, 211)</td>
<td>200</td>
<td>*</td>
</tr>
<tr>
<td>(16, 211)</td>
<td>199</td>
<td>*</td>
</tr>
<tr>
<td>(23, 211)</td>
<td>198</td>
<td>*</td>
</tr>
<tr>
<td>(15, 211)</td>
<td>197</td>
<td>*</td>
</tr>
<tr>
<td>(15, 211)</td>
<td>196</td>
<td>*</td>
</tr>
<tr>
<td>(11, 206)</td>
<td>195</td>
<td>*</td>
</tr>
<tr>
<td>(17, 203)</td>
<td>194</td>
<td>*</td>
</tr>
<tr>
<td>(9, 213)</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>(19, 206)</td>
<td>192</td>
<td></td>
</tr>
<tr>
<td>(10, 208)</td>
<td>191</td>
<td></td>
</tr>
<tr>
<td>(15, 209)</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>(41, 212)</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>(11, 209)</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>(11, 211)</td>
<td>187</td>
<td></td>
</tr>
<tr>
<td>(15, 210)</td>
<td>186</td>
<td></td>
</tr>
<tr>
<td>(12, 204)</td>
<td>185</td>
<td></td>
</tr>
<tr>
<td>(15, 210)</td>
<td>184</td>
<td></td>
</tr>
<tr>
<td>(15, 206)</td>
<td>183</td>
<td></td>
</tr>
<tr>
<td>(14, 202)</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>(25, 208)</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>(21, 208)</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>(13, 208)</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>(14, 205)</td>
<td>178</td>
<td></td>
</tr>
<tr>
<td>(11, 207)</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>(16, 205)</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>(15, 207)</td>
<td>175</td>
<td></td>
</tr>
<tr>
<td>(16, 210)</td>
<td>174</td>
<td></td>
</tr>
<tr>
<td>(12, 202)</td>
<td>173</td>
<td></td>
</tr>
<tr>
<td>(12, 208)</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>(9, 203)</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>(12, 210)</td>
<td>170</td>
<td></td>
</tr>
<tr>
<td>(10, 205)</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>(15, 208)</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>(12, 209)</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>(9, 207)</td>
<td>166</td>
<td></td>
</tr>
<tr>
<td>(7, 206)</td>
<td>165</td>
<td></td>
</tr>
<tr>
<td>(14, 210)</td>
<td>164</td>
<td></td>
</tr>
</tbody>
</table>
Results - Worst

Bootstrap confidence interval of rank
Features (sorted by observed rank)

0 50 100 150 200 250
(3 203) 13
(3 190) 12
(6 209) 11
(3 194) 10
(3 204) 9
(3 200) 8
(4 202) 7
(4 188) 6
(5 203) 5
(4 169) 4
(3 174) 3
(3 201) 2
(4 198) 1
Value of Continuous Biomarkers

- Avoid arbitrary cut-points
- Better risk spectrum
- Provides gray zone
- Increases power/precision
- Fewer biomarkers required to achieve same accuracy
  → 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.

Prognostic Spectrum From Various Models
With Model Chi-square – d.f., and Generalized C Index

Predicted 2-year Disease Recurrence Probability

- PSA+Gleason+Old Stage
  - $X_2$-d.f. = 178, $C$ = 0.77
- PSA+Gleason
  - $X_2$-d.f. = 155, $C$ = 0.75
- PSA
  - $X_2$-d.f. = 92, $C$ = 0.70
- Gleason
  - $X_2$-d.f. = 88, $C$ = 0.68
- New Stage, 6 Levels
  - $X_2$-d.f. = 135, $C$ = 0.73
- New Stage
  - $X_2$-d.f. = 134, $C$ = 0.73
- Old Stage
  - $X_2$-d.f. = 67, $C$ = 0.67

PSA

0.0 0.2 0.4 0.6 0.8
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.
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., 2008, Peek et al., 2007].
- Necessary to unbiasedly validate a high-resolution calibration curve (smooth plot of predicted vs. actual risk of outcome).
The essence of “honesty” or “objectivity” demands that we take into account all the evidence we have, not just some arbitrarily chosen subset of it.

ET Jaynes

*Probability Theory: The Logic of Science*

Failure to use all available information (e.g., conditioning on $X > c$ instead of $X = x$) $\rightarrow$ different researchers ignore different information and get different answers.
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” → more data needed
- Assumes statistician to be provider of utility function
- Sensitivity and specificity are also improper scoring rules
# Sensitivity of Exercise ECG for Diagnosing CAD

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 40</td>
<td>0.56</td>
</tr>
<tr>
<td>40–49</td>
<td>0.65</td>
</tr>
<tr>
<td>50–59</td>
<td>0.74</td>
</tr>
<tr>
<td>≥ 60</td>
<td>0.84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>0.72</td>
</tr>
<tr>
<td>female</td>
<td>0.57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Diseased CAs</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>2</td>
<td>0.68</td>
</tr>
<tr>
<td>3</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Problems Caused by Chopping Continuous Variables

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

<table>
<thead>
<tr>
<th>Range of Delay</th>
<th>Mean Score</th>
<th>Range of Delay</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11</td>
<td>210</td>
<td>0-3.8</td>
<td>220</td>
</tr>
<tr>
<td>11-20</td>
<td>215</td>
<td>3.8-8</td>
<td>219</td>
</tr>
<tr>
<td>21-30</td>
<td>217</td>
<td>8-113</td>
<td>217</td>
</tr>
<tr>
<td>31-40</td>
<td>218</td>
<td>113-170</td>
<td>215</td>
</tr>
<tr>
<td>41-</td>
<td>220</td>
<td>170-</td>
<td>210</td>
</tr>
</tbody>
</table>
There is no apparent relationship between the length of the delay and the examinee’s overall score.

Data from Wainer [2006]
... in almost every study where [finding optimal cut-points] is applied, another cut-point 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 cut-points used in the literature; some of them were solely used because they emerged as the ‘optimal’ cut-point 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 included with 12 different cut-points ... 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]
Summary

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


**Abstract**

This talk will cover some of the bad statistical practices that have crept into biomarker research, including improper use of continuous variables, the use of improper predictive accuracy indexes, and setting the bar too low for demonstrating that biomarker information is new. Some special problems in dealing with higher-dimensional candidate biomarkers will be briefly discussed. Kidney disease poses special problems in biomarker research because commonly used definitions of kidney dysfunction/injury are based on categorizations that violate several statistical and clinical principles. This will also be discussed.