JSM Shortcourse

Statistical Evaluation of Medical Tests and Biomarkers for Classification

July 29, 2007
Introduction

• speakers

• context/applications

• DABS Center website:
  www.fhcrc.org/science/labs/pepe/dabs/
Course Outline

1. Basic ROC Analysis (M. Pepe 8:45–10:15)
2. Study Design Part I (H. Janes 10:30–11:30)
3. Prospective Studies (T. Alonzo 11:30–12:30)

   lunch 12:30–2:00 p.m.

4. Adjusting for Covariates (M. Pepe 2:00–3:00)
5. Study Design Part II (H. Janes 3:00–3:45)
6. Imperfect Reference Tests (T. Alonzo 4:00–4:30)
7. Combining Predictors (M. Pepe 4:30–5:00)
Basic ROC Analysis (M. Pepe)
Notation/Terminology

\( D = \) outcome (disease)

\( Y = \) test result (biomarker)

<table>
<thead>
<tr>
<th></th>
<th>( D = 0 )</th>
<th>( D = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>true negative</td>
<td>false negative</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td>false positive</td>
<td>true positive</td>
</tr>
</tbody>
</table>
Classification Probabilities

TPF = true positive fraction = \(\text{Prob}[Y = 1|D = 1]\) = sensitivity
TPF = false positive fraction = \(\text{Prob}[Y = 1|D = 0]\) = 1-specificity
FNF = false negative fraction = \(\text{Prob}[Y = 0|D = 1]\) = 1-TPF
TNF = true negative fraction = \(\text{Prob}[Y = 1|D = 0]\) = 1-FPF

• (FPF,TPF) is sufficient
• (cost, benefit), 2 dimensions

Ideal test: FPF=0 and TPF = 1
Coronary Artery Surgery Study (CASS)

Coronary Artery Disease

<table>
<thead>
<tr>
<th></th>
<th>$D = 0$</th>
<th>$D = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercise</td>
<td>$Y = 0$</td>
<td>327</td>
</tr>
<tr>
<td>Test</td>
<td>$Y = 1$</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>442</td>
</tr>
</tbody>
</table>

- FPF = 0.26      TPF = 0.80
- Odds Ratio = 11.4, a poor way to summarize accuracy
Classification Probabilities

- FPF = $P(Y = 1|D = 0)$; TPF = $P(Y = 1|D = 1)$

- condition on true status, natural for case-control studies

- question asked by basic scientist, “to what extent does biomarker reflect true status?”

- question asked in decision making “Should I take the test?”; “Should we use the test?”
Continuous biomarker

- *most* biomarkers
- sometimes with “point mass” at a particular value (e.g. 0) or lower limit of detection

Convention

- larger $Y$ more indicative of disease
- formally $P(D = 1|Y)$ increasing in $Y$
Receiver Operating Characteristic (ROC) Curve

- generalizes (FP, TP) to continuous markers
- thresholding positivity rule “Y ≥ C”
- $TPF(c) = P(Y \geq c | D = 1)$
  $FPF(c) = P(Y \geq c | D = 0)$
- $ROC(\cdot) = \{(FPF(c), TPF(c)) ; \ c \in (-\infty, \infty)\}$
Pancreatic cancer biomarkers (Wieand et al 1989)
ROC curves for pancreatic cancer biomarkers

CA 19-9
CA 125

TPF

FPF
Properties/Attributes of the ROC curve

• monotone increasing from (0,0) to (1,1) as threshold decreases from $c = \infty$ to $c = -\infty$

• *ideal* biomarker: has ROC through (0,1)

• invariant to monotone increasing transformations of $Y$, doesn’t depend on scale for $Y$

• puts different biomarkers on a common relevant scale

• shows entire range of possible performance
ROC curves for pancreatic cancer biomarkers

- unequivocally CA-19-9 is the more accurate diagnostic biomarker for pancreatic cancer
Choosing a Threshold (Formal decision theory)

Expected Cost \( (c) = \rho(1 - \text{TPF}(c))C_D^- + (1 - \rho)\text{FPF}(c)C_N^+ \)

Expected Cost \( (t) = \rho(1 - \text{ROC}(t))C_D^- + (1 - \rho)tC_N^+ \)

- minimize cost at the threshold that corresponds to

\[
\frac{\delta\text{ROC}(t)}{\delta t} = \frac{1 - \rho}{\rho} \frac{C_N^+}{C_D^-}
\]
A pancreatic cancer biomarker in 90 cases and 51 controls. Shown are (a) raw biomarker distributions; (b) the empirical ROC curve; (c) skill plots, and (d) cost plots. In (c) and (d) we assume three different values for standardized false positive cost values, $\theta = 0.1, 0.5, 0.9$, and we assume the sample prevalence is the population disease prevalence. $\theta = C_N^+ 1 - \theta = C_D^-$.
Choosing a Threshold (common informal practice)

- fix maximum tolerated FPF, e.g. must be very low in disease screening of healthy populations
- FPF $= f_0 \rightarrow threshold = 1 - f_0$ quantile of $Y$ in controls
- fix minimum TPF, e.g. must often be high in diagnostic settings
- FPF $= t_0 \rightarrow threshold = 1 - t_0$ quantile of $Y$ in cases
Focused ROC analysis

- specific regions of FPF and/or TPF axis are often of interest
- e.g., low FPF in screening
  e.g., high TPF in diagnostics

- plot relevant part of the ROC curve
Demonstration of Software

- www.fhcrc.org/science/labs/pepe/dabs/
- under ‘software’ we use Stata ‘basic ROC analysis’
- under ‘datasets’ use pancreatic or ovarian cancer data
- download and enter Stata
- install ‘pcvsuite’
- use ‘help roccurve’ to see how to use the roc estimation command;
  use ’help comproc’ to see how to use roc comparison command
- e.g.‘roccurve D Y1, roct(0.2) ci’ provides the empirical ROC
Empirical ROC curve for the pancreatic cancer marker CA-19-9 with confidence interval at FPF=0.2
Empirical ROC Curve

• \( \hat{TPF}(c) = \) proportion of cases with \( Y \geq c \)
  \( \hat{FPF}(c) = \) proportion of controls with \( Y \geq c \)

• plot \( \hat{TPF}(c) \) versus \( \hat{FPF}(c) \) \( c \in (-\infty, \infty) \)

• a step function: step down at a case value
  step left at a control value

• depends only on ranks of data, not the raw values
Notation

\( \bar{D} \): controls, \( n_{\bar{D}} \) of them

\( D \): cases, \( n_D \) of them

\( \hat{S}_D(y) = \hat{P}(Y_D \geq y) \)

\( t \equiv \text{FPF} \)

\( \hat{S}_{\bar{D}}^{-1}(t) = y \) such that \( \hat{S}_{\bar{D}}(y) = t \)

=\ threshold\ corresponding\ to\ \text{FPF} = t

\( \hat{\text{ROC}}_e(t) = \) empirical ROC at \( \text{FPF} = t \)

Estimation Steps

\[
t \xrightarrow{n_{\bar{D}}} \left( \begin{array}{c}
\text{estimated \ threshold} \\
= \hat{S}_{\bar{D}}^{-1}(t)
\end{array} \right) \xrightarrow{n_D} \left( \begin{array}{c}
R\hat{\text{OC}}_e(t) \\
= \hat{S}_D \left( \hat{S}_{\bar{D}}^{-1}(t) \right)
\end{array} \right).
\]
Result

$$\text{var}(\hat{\text{ROC}}_e(t)) = \frac{\text{ROC}(t)\{1 - \text{ROC}(t)\}}{n_D} + (\text{ROC}'(t))^2 \frac{t(1 - t)}{n_{\bar{D}}}$$

- Second component due to estimating threshold. 
  $\text{ROC}'(t) = \text{derivative of ROC at } t$. 
  larger $n_{\bar{D}}$ needed when slope of ROC is high.

- First component due to estimating TPF at threshold. Larger $n_D$ required when $\text{ROC}(t) \approx 0.5$
Normal Theory Confidence Intervals

\[
\text{CI}(\alpha) = (\hat{\text{ROC}}_e(t) - Z^{\alpha/2} \sqrt{\text{var}}, \hat{\text{ROC}}_e(t) + Z^{\alpha/2} \sqrt{\text{var}})
\]

where \( Z^{\alpha/2} = \) normal deviate

Bootstrap Confidence Intervals

- resample with replacement \( n_D \) cases and \( n_D \) controls, calculate \( \hat{\text{ROC}}(t) \)
- repeat \( B \) times: \( \{\hat{\text{ROC}}^1(t), \ldots, \hat{\text{ROC}}^B(t)\} \)
- select \( \alpha/2 \) and \( 1 - \alpha/2 \) quantiles, \( (r^L, r^H) \)

\[
\text{CI}(\alpha) = (r^L, r^H)
\]
• in Stata ‘roccurve d y1, roct(0.2) ci’ yields confidence interval at $t = 0.2$ using bootstrap technique

• default 95% confidence level, ‘cilevel’ to change default $B=1000$ bootstrap samples, ‘nsamp’ to change default assumes $n_D$ and $n_{\bar{D}}$ fixed, ‘roccsamp’ to accommodate cohort rather than case-control design

• if several observations come from the same test unit (e.g., person) use the ‘cluster’ option in bootstrap resampling

• bootstrap confidence intervals vary because of random resampling
Example

- Marker sought for screening for ovarian cancer
- allow no more than 2% false positives
- data on two markers for 1000 controls and 200 cases
- What % of breast cancers detected with Y?
- Current marker, detects 30% of cancers. Do either of the new markers detect >30% of cancers?
- Analyze ovarian-markers dataset: $Y_1$ and $Y_2$
• Y1: $ROC(0.02) = 0.155 \ (0.100, 0.218)$

• Y2: $ROC(0.02) = 0.470 \ (0.375, 0.558)$
Compare Markers

- $\text{ROC}_1(t)$ versus $\text{ROC}_2(t)$ at a $t$ of interest
- ovarian cancer acceptable FPF fixed at $t = 2\%$
- $\widehat{\text{ROC}}_1(t) - \widehat{\text{ROC}}_2(t) =$ difference in % of diseased detected when both markers allow FPF=$2\%$
  
  - $0.470 - 0.155 = 0.315$
- $95\%$ CI based on bootstrap is $(0.215, 0.425)$
- conclude $Y_2$ detects 21.5 to 42.5% more cancers; 
  
  $p$-value $< 0.001$
Area Under the ROC Curve (AUC)

\[ \text{AUC} = \int_0^1 \text{ROC}(t) \, dt = \text{average (TPF)} \]

- averaging uniformly across \( FPF \in (0, 1) \)
- commonly used summary ROC index

- \textit{ideal} test: AUC=1.0
- \textit{useless} test: AUC=0.5
Another Interpretation

• $\text{AUC} = \text{Prob}(Y_D \geq Y_{\bar{D}})$
  = probability of correct ordering for random case and control pair

• a clinically relevant measure of performance?

• NO!
Example (ovarian cancer dataset)

AUC for $Y_1 = 0.770$
AUC for $Y_2 = 0.922$
$p$-value $< 0.001$; Wald test

ROC(0.02) for $Y_1 = 0.155$
ROC (0.02) for $Y_2 = 0.470$
$p$-value $<0.001$

- conclusions about ROC(0.02) are more clinically important than those about AUC
The Partial AUC

\[ pAUC(t) = \int_0^t ROC(f) df \]

\[ pAUC(t)/t = TPF \text{ averaged over } FPF \in (0, t) \]

- used to compare two ROC curves over a clinically relevant range of FPFs
- a compromise between AUC and ROC(t)
### Ovarian Cancer Data

<table>
<thead>
<tr>
<th></th>
<th>ROC(_1)(0.02) − ROC(_2)(0.02)</th>
<th>(pAUC(_1)(0.02) − pAUC(_2)(0.02))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>estimate</strong></td>
<td>0.315</td>
<td>0.271</td>
</tr>
<tr>
<td><strong>95% CI</strong></td>
<td>(0.215, 0.425)</td>
<td>(0.179, 0.365)</td>
</tr>
<tr>
<td><strong>p-value</strong></td>
<td>&lt; .001</td>
<td>&lt; .001</td>
</tr>
<tr>
<td><strong>uses</strong></td>
<td>less data</td>
<td>more data</td>
</tr>
<tr>
<td><strong>power for ordered ROC curves</strong></td>
<td>less</td>
<td>more</td>
</tr>
<tr>
<td><strong>interpretation</strong></td>
<td>excellent</td>
<td>okay</td>
</tr>
</tbody>
</table>

29
Inference for $\text{ROC}^{-1}$

- Fix $\text{TPF} = v$. What is corresponding $\text{FPF}$?
- $\text{ROC}^{-1}(v)$

- estimation, CI etc. for ROC can be achieved with methods already described.

- Need to redefine disease and biomarker:
  $D^* = 1 - D$, reverse case-control status
  $Y^* = -Y$, reverse biomarker

- $\text{FPF}^* = P[Y^* > c|D^* = 0] = P[-Y > c|D = 1] = 1 - \text{TPF}$
  $\text{TPF}^* = P[Y^* > c|D^* = 1] = P[-Y > c|D = 0] = 1 - \text{FPF}$

- Fix $\text{FPF}^* = 1 - v$; $\text{ROC}^{-1}(v)$ is corresponding $1 - \text{TPF}^*$
Ovarian Cancer Data

- Find FPF corresponding to $\text{TPF} = 80\%$ for $Y_2$

- $\text{FPF}^* = 0.2$ yields $\text{TPF}^* = 0.864$
  
  i.e., $\text{TPF} = 0.8 \Rightarrow \text{FPF} = 0.136$

- $95\%$ CI for $\text{TPF}^* = (0.814, 0.910)$
  
  i.e., $95\%$ CI for $\text{ROC}^{-1}(0.8)$ i.e. for FPFis $(.090, .186)$
Ovarian Cancer Data

marker: y2s

TPF

FPF
Compare Markers

- with respect to $\text{ROC}^{-1}(0.8)$
- set $\text{FPF}^* = 0.2$ and calculate corresponding $\text{TPF}^*$
  - $\text{TPF}_1^* = 0.599$ i.e., $\text{ROC}^{-1}(0.8) = 0.401$
  - $\text{TPF}_2^* = 0.864$ i.e., $\text{ROC}^{-1}(0.8) = 0.136$

$p < 0.001$ $p < 0.001$
Binormal ROC Curve

- the classic ROC curve

- $Y_D \sim N(\mu_D, \sigma_D^2)$  $Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2)$

  implies

  $$\text{ROC}(t) = \Phi(a + b\Phi^{-1}(t))$$

  where

  $$\Phi(\cdot) = \text{normal distribution cdf}$$
  $$= \text{Prob(standard normal variable } \leq \cdot)$$

- $a =$‘intercept’; $b =$‘slope’

  $$a = (\mu_D - \mu_{\bar{D}})/\sigma_D \quad b = \sigma_{\bar{D}}/\sigma_D$$
• $\text{ROC}(t) = \Phi(a + b\Phi^{-1}(t))$

• $a$ determines TPF at $t = 0.5$
  
  $b$ determines ‘slope’

• a classic family of ROC curves
Underlying Binormal Assumption

- for some unspecified transformation $g$
  
  $$g(Y_D \bar{D}) \sim \text{Normal} \quad \text{and} \quad g(Y_D) \sim \text{Normal}$$

- raw data does not have to be normally distributed

- once $a$ and $b$ are specified, smooth ROC curve can be obtained
Ovarian Cancer Data

- $Y_1$: ROC(0.02) = 0.175 (0.113, 0.251)
- $Y_2$: ROC(0.02) = 0.392 (0.225, 0.515)
Fitting over a Subinterval

- $Y_1$: $\text{ROC}(0.02) = 0.136$ (0.081, 0.187)
- $Y_2$: $\text{ROC}(0.02) = 0.453$ (0.371, 0.526)
How is Binormal Fitting Done?

- Choose $t_1, t_2, \ldots, t_{np}$
- Find corresponding thresholds $y_1, y_2, \ldots, y_{np}$
- $\text{ROC}(t_k) = \Phi(a + b\Phi^{-1}(t_k))$
- That is, $P(Y_D > y_k) = \Phi(a + b\Phi^{-1}(t_k))$
- For each case generate $U_k = I(Y > y_k)$
- probit regression for $U_k$ with covariate $\Phi^{-1}(t_k)$
  fits the model $P(U_k = 1) = P(Y_D > y_k) = \Phi(a + b\Phi^{-1}(t_k))$
Binormal versus Empirical ROC curve

- binormal is smooth: one-one function esthetically pleasing; same as true ROC
- binormal is statistically more efficient
- binormal depends on assumptions, albeit weak assumptions

Common Practical uses of Binormal Curve

- in radiology where markers are ordinal ratings of images, not continuous markers
- sample size calculations, have to assume something about form of data expected
- ROC regression analysis
Percentile Values

\[ pv(Y) = \text{proportion of controls with marker values} < Y \]

- a natural standardization of \( Y \)
- puts different markers on a common scale
- for controls \( pv(Y) \sim \text{uniform} (0,1) \)
- for cases a good marker will tend to have high \( pv(Y) \)
**Another Interpretation for ROC Curve**

**Result**

\[
\text{ROC}(t) = \text{Prob}[1 - t \leq pv(Y_D)] \\
= \text{Prob}[1 - pv(Y_D) \leq t]
\]

**Proof**

\[
\text{Prob}[pv(Y_D) \geq 1 - t] = \text{Prob}[Y_D \geq \text{threshold for FPF} = t]
\]

- ROC is the cumulative distribution of standardized marker values for cases
- controls used only to standardize the marker
**Figure**: Case cumulative distributions of standardized CA-19-9
Result

\[
\text{AUC} = \text{average } (pv)
\]
\[
p\text{AUC}(t) = \text{average } (\max(pv - (1 - t), 0))
\]
\[
\text{ROC}(t) = \text{Prob}(pv(Y_D) \geq 1 - t)
\]

- comproc uses this formulation to compare markers
- roccurve uses percentile value formulaiton to estimate ROC curves
Percentile Value Standardization

- traditional statistical elements: standardization, usual statistical methods
- easy programming
- concepts such as covariate adjustment follow
Percentile Value Calculation

- empirical: \( pv(y) = \text{proportion of controls with } Y_\bar{D} < y \)

- normal: approximate distribution of \( Y_\bar{D} \) with \( Y_\bar{D} \sim N(\mu_\bar{D}, \sigma^2_\bar{D}) \)
estimate mean and variance from control data
calculate

\[
 pv(y) = \Phi((y - \mu_\bar{D})/\sigma_\bar{D})
\]
commonly called a \( Z\)-score
Empirical ROC estimates using normal theory percentile scores

- $Y_1$: $\text{ROC}(0.02) = 0.160$ (0.110, 0.220)
- $Y_2$: $\text{ROC}(0.02) = 0.465$ (0.375, 0.538)
Tied marker values

- When case and control values are tied, ROC has simultaneous horizontal and vertical jumps

- AUC and pAUC(t) traditionally calculated with trapezoidal rule

- can be achieved by calculating AUC and pAUC(t) as averages of tie-corrected percentile values

- tie-corrected \( pv(y) = P(Y_D < y) + 0.5P(Y_D = y) \)
Summary

• motivation for (TPF, FPF) and ROC curve

• estimation of ROC with nonparametric (empirical) or semiparametric (binormal) methods

• comparison statistics: \[ \hat{\text{ROC}}(t) \]
  \[ \hat{\text{ROC}}^{-1}(v) \]
  partial AUC(t)
  AUC, but not encouraged

• percentile value standardization
Bibliography


Study Design Part I
Outline

• phases of research for biomarker development

• sample size calculations for early phase studies
Phases of Research for Therapeutic Drugs

phase 1: toxicity/dose finding
phase 2: biologic response
phase 3: clinical response
phase 4: post-marketing response

- leads to reasonably rigorous and efficient research
- early phases weed out drugs with little or toxic effects
- scientific objectives at each phase make sense in the development process
- data from early phase studies are used to design later studies
- appropriate study designs follow naturally
Phases of Research for Biomarker Development

- Early Detection Research Network (EDRN): 5-phase structure for development of cancer biomarkers (Pepe et al. 2001)
- Zhou et al. (2002) Phases of design for diagnostic accuracy studies
- Baker et al. (2006) Evaluating markers for the early detection of cancer: overview of study designs and methods
<table>
<thead>
<tr>
<th>#</th>
<th>Phase</th>
<th>Objective</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Preclinical Exploratory</td>
<td>promising directions identified</td>
<td>diverse and convenient cases and controls</td>
</tr>
<tr>
<td>2</td>
<td>Clinical Assay and Validation</td>
<td>clinical assay detects established disease</td>
<td>population based, cases with disease, controls without disease</td>
</tr>
<tr>
<td>3</td>
<td>Retrospective Longitudinal</td>
<td>biomarker detects disease <em>early</em> before it becomes clinical</td>
<td>case-control study nested in a longitudinal cohort</td>
</tr>
<tr>
<td>4</td>
<td>Prospective Screening</td>
<td>extent and characteristics of disease detected by the test and the false referral rate are identified</td>
<td>Cross-sectional cohort of <em>people</em></td>
</tr>
<tr>
<td>5</td>
<td>Cancer Control</td>
<td>impact of screening on reducing the burden of disease on the population is quantified</td>
<td>randomized trial (ideally)</td>
</tr>
</tbody>
</table>
Phase 1 (Exploratory)

Objectives

- to identify promising tests and settings
- assess test reproducibility

Design: case-control study with convenience sampling

Principle

- diversity in selecting cases, controls, settings, operating parameters, testers, etc.
- hypothesis generating
Phase 2 (Validation)

Objectives

(i) to determine if the test detects established disease with minimally acceptable accuracy

(ii) to define criteria for test ‘positivity’

(iii) to compare the accuracy of the new test with the standard of practice

(iv) to assess the impact of factors that affect the test result in non-diseased subjects

(v) to assess factors that affect test sensitivity or ROC curve
Phase 2, cont’d

Design: case-control study

- random selection of cases and controls
- matching should be carefully considered (more later)
- cases with the range of disease that should be detected

Procedures/Evaluations

- rigorous definitions of protocol, setting, selection criteria
Phase 3 (Validation for Early Detection)

Design: case-control study nested within longitudinal cohort

- obtain samples prior to disease diagnosis
- matching should be carefully considered

Objectives

(i) to determine if the test detects disease prior to diagnosis with minimally acceptable accuracy

(ii) to define criteria for test ‘positivity’

(iii) to compare the accuracy of the new test with the standard of practice

(iv) to assess the impact of factors that affect the test result in non-diseased subjects

(v) to assess factors that affect test sensitivity or ROC curve
Phase 4 (Cohort Study)

• prospective study: apply test to subjects with unknown disease status
• testing often determines patient care
• subjects who test negative may not have disease status ascertained

Objectives:

• determine operating characteristics of test as diagnostic tool
• determine extent and characteristics of disease detected
• will subjects comply?
• will subjects go for treatment?
• how much does this cost?
Sample Size Calculations for Phase 2 and 3 Studies
Phase 2 and 3 Studies

Objective:

• to determine if the test detects disease with minimally acceptable accuracy

Design:

• case-control study

Measures of Test Accuracy:

• TPF/FPF for binary marker

• ROC curve for continuous marker
Sample Size Calculations for Binary Markers
Scientific Question

• Does the test have minimally acceptable accuracy?

• identify (FPF, TPF) that are minimally acceptable
  \[ FPF \leq FPF_0 \] and \[ TPF \geq TPF_0 \]

• specified by investigators

• depend on the context
  – costs/benefits associated with false/true positive classifications
  – given available resources, population where test is applied
Determining whether the test has minimally acceptable performance

• calculate observed (FPF, TPF) and confidence interval
  – joint confidence region for (FPF, TPF): product of individual confidence intervals (rectangular region)
  – exact binomial or asymptotic normal confidence intervals

• positive conclusion if $\text{FPF} \leq \text{FPF}_0$ and $\text{TPF} \geq \text{TPF}_0$ with high confidence
  – if joint confidence interval lies entirely within the region of acceptable values
Figure 8.2  A one-sided rectangular confidence region for (FPF, TPF) of the Exercise Stress Test calculated from the CASS data. The classification probabilities meet the minimal criteria: $\text{FPF} \leq .35$ and $\text{TPF} \geq .70$. The points indicated with asterisks denote $\text{FPF}_{\alpha^*}^U$ and $\text{TPF}_{\alpha^*}^L$, the upper and lower $\alpha^* = 1 - \sqrt{1 - \alpha}$ confidence limits for FPF and TPF, respectively.

From The Statistical Evaluation of Medical Tests for Classification and Prediction by Margaret S. Pepe, Ph.D., Oxford University Press, 2003
Sample Size Formulae Based on Asymptotic Variances

For testing

\[ H_0 : (TPF \leq TPF_0 \text{ or } FPF \geq FPF_0) \]

in order to achieve type-I error \( \alpha \) and power \( 1 - \beta \) at \((FPF_1, TPF_1)\), we require

\[
\begin{align*}
 n_D &= \frac{(Z^{1-\alpha^*} \sqrt{TPF_0(1-TPF_0)} + Z^{1-\beta^*} \sqrt{TPF_1(1-TPF_1)})^2}{(TPF_1 - TPF_0)^2} \\
n_{\tilde{D}} &= \frac{(Z^{1-\alpha^*} \sqrt{FPF_0(1-FPF_0)} + Z^{1-\beta^*} \sqrt{FPF_1(1-FPF_1)})^2}{(FPF_1 - FPF_0)^2}
\end{align*}
\]

where \( \alpha^* = 1 - \sqrt{1-\alpha} \) and \( \beta^* = 1 - \sqrt{1-\beta} \), \( Z^{1-\alpha^*} = \Phi^{-1}(1 - \alpha^*) \), \( Z^{1-\beta^*} = \Phi^{-1}(1 - \beta^*) \).

- based on asymptotic variances
- confidence rectangle: product of two one-sided confidence intervals
A Note on the Choice of Null Hypothesis

Testing $H_0: TPF = FPF$ usually not of interest

- corresponds to useless test

- null hypothesis should be that the test is less than minimally accurate
  - reject null $\implies$ test is at least minimally accurate

- $TPF = FPF$ is usually far less than minimally accurate
  - eg use of a cancer screening test with $TPF = FPF = 0.5$ would be irresponsible
Example: diagnostic test for chlamydia

- new urinary test for chlamydia
- hope it will be 95% specific, 90% sensitive
- must show it is at least 80% specific, 75% sensitive
  \[(FPF_0, TPF_0) = (0.20, 0.75)\]
  \[(FPF_1, TPF_1) = (0.05, 0.90)\]
- conclusions based on confidence rectangle
- formulae for 90% power and 10% type-I error indicate 64 cases, 46 controls required
Sample Size Calculations Using Exact Methods

- simulate a large number of datasets of size \((n_{D_0}, n_{\bar{D}_0})\) with 
  \[ Y_D \sim Bin(TPF_1), \ Y_{\bar{D}} \sim Bin(FPF_1) \]

- in each dataset, calculate exact rectangular confidence interval of size \(\alpha\)

- calculate Power = proportion of times that confidence interval rules out null hypothesis,

  \[ H_0 : (TPF \leq TPF_0 \text{ or } FPF \geq FPF_0) \]

- use asymptotic sample sizes as starting points for the simulations
Stata Program

- `screensize` package
- `scrsise` function

- specify:
  - null \((\text{FPF}_0, \text{TPF}_0)\) and alternative \((\text{FPF}_1, \text{TPF}_1)\)
  - type-I error, \(\alpha\)
  - \((n_{D_0}, n_{\bar{D}_0})\)

- performs simulations and returns power
Chlamydia example, cont’d

- simulations with \((n_{D_0}, n_{\bar{D}_0}) = (64, 46)\) indicate 88% power
- simulations show 91% power achieved with 70 cases, 50 controls
Sample Size Calculations for Continuous Markers
Scientific Question

- for some threshold $c$ does the dichotomized test $I[Y > c]$ have acceptable performance?
- ie can test operate at or above minimally acceptable TPF and FPF?
Figure 8.3  ROC curves for two (hypothetical) tests. The upper one meets the minimally acceptable criterion in that it can attain operating points that exceed \((FPF_0, TPF_0)\), whereas the lower one does not.

From *The Statistical Evaluation of Medical Tests for Classification and Prediction* by Margaret S. Pepe, Ph.D., Oxford University Press, 2003
Determining whether the test has minimally acceptable performance

- choose minimally acceptable \((FPF_0, TPF_0)\)
- does the TPF of the test exceed \(TPF_0\) at the threshold corresponding to \(FPF_0\)?
  - i.e. is \(ROC(FPF_0) \geq TPF_0\)
- calculate observed \(ROC(FPF_0)\) and confidence interval
- positive conclusion if \(ROC(FPF_0) \geq TPF_0\) with high confidence
  - if lower limit of confidence interval for \(ROC(FPF_0)\) exceeds \(TPF_0\)
Figure 8.4  A test of the null hypothesis that at the threshold corresponding to FPF$\_0=0.2$, the TPF does not exceed 0.6 for the CA 19-9 marker of pancreatic cancer. Shown is the lower 95% confidence limit for ROC(0.2) using data from Wieand et al (1989).

From The Statistical Evaluation of Medical Tests for Classification and Prediction
by Margaret S. Pepe, Ph.D., Oxford University Press, 2003
Sample Size Formula Based on Asymptotic Variance

For testing

\[ H_0 : \text{ROC}(\text{FPF}_0) \leq \text{TPF}_0 \]

in order to achieve type-I error \( \alpha \) and power \( 1 - \beta \) at \( \text{TPF}_1 = \text{ROC}(\text{FPF}_0) \) we require

\[
n_D = \frac{\left\{ \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) \right\}^2}{(\text{TPF}_0 - \text{TPF}_1)^2} \cdot V_1
\]

where \( V_1 = n_D \text{ Var}(\hat{\text{ROC}}_e(\text{FPF}_0)) \)

- based on lower limit of one-sided confidence interval for \( \text{ROC}(\text{FPF}_0) \)
\[ V_1 = n_D \text{Var}(\widehat{\text{ROC}}_e(t)) = \text{ROC}(t)(1 - \text{ROC}(t)) + \lambda \left\{ \frac{\partial \text{ROC}(t)}{\partial t} \right\}^2 t(1 - t) \]

- \( \lambda = \frac{n_D}{n_\bar{D}} \), the case-control ratio

- \( \lambda_{opt} = \left\{ \frac{\partial \text{ROC}(t)}{\partial t} \right\}^{-1} \sqrt{\frac{\text{ROC}(t)(1 - \text{ROC}(t))}{t(1 - t)}} \), (Janes and Pepe, 2006)
  - minimizes asymptotic variance

- Input:
  - \( t = \text{FPF}_0 \)
  - \( \text{ROC}(t) = \text{TPF}_1 \)
  - \( \frac{\partial \text{ROC}(t)}{\partial t} \) at \( t = \text{FPF}_0 \)
Estimating ROC Slope

- based on \( \frac{\partial \text{ROC}(t)}{\partial t} = \frac{\partial S_D(S_D^{-1}(t))}{\partial t} = \frac{f_D(S_D^{-1}(t))}{f_D(S_D^{-1}(t))} = \text{LR}(S_D^{-1}(t)) \)
  - ratio of kernel density estimates, evaluated at \( S_D^{-1}(t) \)

- based on parametric model for the ROC curve
  - eg \( \text{ROC}(t) = \Phi(a + b \Phi^{-1}(t)) \) implies \( \frac{\partial \text{ROC}(t)}{\partial t} = b \frac{\phi(a + b \Phi^{-1}(t))}{\phi(\Phi^{-1}(t))} \)
Example: Barrett’s Esophagus

- diagnostic marker for adenocarcinoma
- $\text{FPF}_0 = 0.10$ because of invasive work-up.
- $\text{TPF}_1 = 0.37$; $\text{TPF}_0 = 0.25$ minimally useful
  
  $\frac{\partial \text{ROC}(t)}{\partial t} = 0.96$ at $\text{FPF}_0$ from pilot data
  
  - estimated using ratio of kernel density estimates
- 90% power, 5% type 1 error
- $\lambda_{opt} = 1.68$, $n_D = 235$, $n_{\bar{D}} = 140$
Sample Size Calculations Using Simulations

- simulate a large number of datasets of size \((n_{D_0}, n_{\bar{D}_0})\) under alternative 
  \(\text{ROC}(\text{FPF}_0) = \text{TPF}_1\)
- assume binormal ROC curve
- in each dataset, calculate confidence interval for \(\text{ROC}(\text{FPF}_0)\) of size \(\alpha\)
- calculate Power = proportion of times that confidence interval rules out null hypothesis,
  
  \[ H_0 : \text{ROC}(\text{FPF}_0) \leq \text{TPF}_0 \]

- use asymptotic sample sizes as starting points for the simulations

- Stata function \texttt{rocsiz}e
Design based on ROC($t$) or AUC?

- ROC($t$) = proportion of disease detected allowing $t$ false referrals

- AUC = $P(Y_D > Y_{\bar{D}})$
  - cute but not clinically meaningful
  - how to specify null value AUC$_0$?

- design based on AUC:
  - does not require specifying $t$
  - usually requires smaller sample sizes
  - but study based on AUC generally not powered to estimate ROC($t$)
Design based on $\text{ROC}^{-1}(v)$

- when control of TPF is required
  - e.g. diagnostic test
- choose minimally acceptable $(\text{FPF}_0, \text{TPF}_0)$
- is $\text{ROC}^{-1}(\text{TPF}_0) \leq \text{FPF}_0$?
- calculate observed $\text{ROC}^{-1}(\text{TPF}_0)$ and confidence interval
- positive conclusion if $\text{ROC}^{-1}(\text{TPF}_0) \leq \text{FPF}_0$ with high confidence
  - if upper limit of confidence interval for $\text{ROC}^{-1}(\text{TPF}_0)$ is less than $\text{FPF}_0$
Example: Breast cancer biomarker

- intended to reduce unnecessary biopsies
- $\text{TPF}_0 = 0.98$, $\text{FPF}_0 = 0.75$
- $\text{FPF}_1 = 0.50$
- 90% power, 5% type 1 error
- require 300 cases and 100 controls
Summary: Study Design Part I

- Biomarker development is a process. Know where you are in the process to identify study questions and study design.

- Phase 1: exploratory

- Phase 2: retrospective validation

- Phase 3: validation for early detection

- Phase 4: prospective validation

- Phase 5: population impact
Summary, cont’d

Sample Size Calculations for Phase 2/3 Studies:

- determine parameter(s) of interest
  - (FPF, TPF) for binary test
  - \( \text{ROC}(t)/\text{ROC}^{-1}(v) \) for continuous test

- characterize minimally useful test (null hypothesis)

- identify desired performance (alternative hypothesis)

- for continuous test, calculate case-control ratio

- calculate starting points for sample sizes using asymptotic formulae

- validate using simulations
References


Selected Verification

Todd A. Alonzo, PhD
University of Southern California
talonzo@childrensoncologygroup.org

July 29, 2007
Outline

1. What is verification bias?

2. Verification bias correction
   • Binary tests
   • Continuous tests

3. Design of selected verification studies
Setting

• Assess diagnostic accuracy of a screening test
• Ideally all study subjects receive screening test and gold standard (GS) test
Verification bias

• may not be possible to obtain disease status for all study subjects
  – too expensive (cost or resources)
  – not ethical (biopsy)
• studies designed so result of test under investigation affects disease verification process
• selective sampling can lead to biased estimates of accuracy - “verification bias” or “work-up bias”
Example of Selected Verification

- 6,691 men received PSA & DRE screening for prostate cancer 1995-2001
- Biopsy recommended if PSA>2.5ng/ml or DRE+
- 705 (11%) received biopsy
  - 70.2% w/ PSA>2.5 or DRE+

Punglia et al. (2003) NEJM
Selected verification is common

• 44% of 16 studies of flow rate testing in urology (Patel et al, 2003)
• 36% of 54 pediatric diagnostic tests (Bates et al, 1993)
• 82% of 62 studies of Pap tests (Fahey et al, 1995)
• 94% of 33 studies of exercise tests for diagnosing CHD (Philbrick et al., 1980)
# Selected Verification

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>Observe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D+</td>
<td>D-</td>
</tr>
<tr>
<td>Test+</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>Test-</td>
<td>20</td>
<td>810</td>
</tr>
</tbody>
</table>

TPR = 80/100 = 0.8
FPR = 90/810 = 0.1
PPV = 80/170 = 0.47
NPV = 810/830 = 0.98

TPR = 80/82 = 0.98
FPR = 90/171 = 0.53
PPV = 80/170 = 0.47
NPV = 81/83 = 0.98
Effect of Selected Verification on Accuracy Estimates

• If verification function of only test results, unadjusted PPV and NPV are unbiased estimates of true PPV and NPV

• If test+ more likely to have GS performed,
  – Unadjusted TPF overestimates true TPF
  – Unadjusted FPF overestimates true FPF

• Incorrect conclusions can be drawn if selected verification not properly accounted for
Missing Data Mechanism

• Missing completely at random (MCAR)
  – Disease verification independent of observed & unobserved data
  – Verification in simple random sample
  – Complete case estimates unbiased

• Missing at random (MAR)
  – Verification depends on observed data
  – \( P(D \mid Y) = P(D \mid Y, V+) \)
  – \( P(V \mid Y) = P(V \mid Y, D) \)
  – Violated if verification depends on unobserved data
    • E.g., symptoms of disease, family history
Correcting for Selected Verification

• Using Bayes’ Rule (Begg & Greenes, 1983)

\[
TPF = P(Y^+ | D^+) = \frac{P(D^+ | Y^+)P(Y^+)}{P(D^+)}
\]

\[
= \frac{P(D^+ | Y^+)P(Y^+)}{P(D^+ | Y^+)P(Y^+) + P(D^+ | Y^-)P(Y^-)}
\]

\[
FPF = P(Y^+ | D^-) = \frac{P(D^- | Y^+)P(Y^+)}{P(D^- | Y^+)P(Y^+) + P(D^- | Y^-)P(Y^-)}
\]

• MAR assumption – PPV, NPV can be estimated using those verified

• Equivalent to inverse probability weighted estimates - divide observed data by estimate of P(V=1|Y)
Correcting for Selected Verification

GS

\[
\begin{array}{ccc}
Y & + & - & ? \\
\hline
 & a & b & e \\
 & c & d & f \\
\end{array}
\]

\[
TPF_{adj} = \frac{n_{Y^+} \ [a/(a+b)]}{n_{Y^+} \ [a/(a+b)] + n_{Y^-} \ [c/(c+d)]}
\]

\[
FPF_{adj} = \frac{n_{Y^-} \ [b/(c+d)]}{n_{Y^-} \ [d/(c+d)] + n_{Y^+} \ [b/(a+b)]}
\]

\[
p\hat{rev}_{adj} = \frac{n_{Y^+} \ [a/(a+b)] + n_{Y^-} \ [c/(c+d)]}{n}
\]
Example: Syphillis Testing

- Motivation – rapid syphilis screening tests needed in rural areas
- Goal – assess accuracy of rapid test (RT)
- Study pop’n – 770 women in rural Peru
- RT+: all tested with GS
- RT–: 38% tested with GS

Garcia et al (2001)
Syphillis Testing

GS

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>-</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>13</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>-</td>
<td>6</td>
<td>275</td>
<td>468</td>
</tr>
</tbody>
</table>

\[ TPF_{\text{unadj}} = \frac{13}{19} = 68.4\% \]

\[ FPF_{\text{unadj}} = \frac{6}{281} = 2.1\% \]

\[ PPV_{\text{unadj}} = \frac{13}{19} = 6.8\% \]

\[ NPV_{\text{unadj}} = \frac{275}{283} = 97.2\% \]
# Syphilis Testing

**GS**

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>-</th>
<th>?</th>
<th>(\text{unadj}) Y+ Y+ Y-</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>13</td>
<td>8</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>-</td>
<td>6</td>
<td>275</td>
<td>468</td>
<td>749</td>
</tr>
</tbody>
</table>

\[
\hat{T}_{\text{PR}}_{\text{unadj}} = \frac{13}{19} = 68.4\%
\]

\[
\hat{T}_{\text{PR}}_{\text{adj}} = \frac{n_{Y+} [a/(a+b)]}{n_{Y+} [a/(a+b)] + n_{Y-} [c/(c+d)]}
\]

\[
= \frac{21 [13/21]}{21 [13/21] + 749 [6/281]}
\]

\[
= 44.8\%
\]
# Syphilis Testing

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prev</td>
<td>6.3%</td>
<td>3.8%</td>
</tr>
<tr>
<td>TPF</td>
<td>68.4%</td>
<td>44.8%</td>
</tr>
<tr>
<td>FPF</td>
<td>2.8%</td>
<td>1.1%</td>
</tr>
<tr>
<td>PPV</td>
<td>61.9%</td>
<td>61.9%</td>
</tr>
<tr>
<td>NPV</td>
<td>97.9%</td>
<td>97.9%</td>
</tr>
</tbody>
</table>
Verification Bias CIs

- CIs wider than if all were verified
- let $\beta = \logit (\text{TPF}_{\text{adj}}) = \log \left( \frac{\text{TPF}_{\text{adj}}}{1 - \text{TPF}_{\text{adj}}} \right)$

- CI for $\beta$: $(\beta_L, \beta_U) = \hat{\beta} \pm Z_{1-\alpha/2} \left( \text{var}(\hat{\beta}) \right)^{1/2}$

- $\text{var}(\hat{\beta}) = \frac{1}{N} \left( \frac{1}{\tau(1-\tau)} + \frac{1 - PPV}{PPV P_1^V \tau} + \frac{1 - NPV}{NPV P_0^V (1-\tau)} \right)$

$\tau = P(Y = 1)$, $P_1^V = P(V = 1 \mid Y = 1)$, $P_0^V = P(V = 1 \mid Y = 0)$

$\text{var}(\hat{\beta})$: $\frac{n}{n_{Y+}} + \frac{b}{a(a+b)} + \frac{d}{c(d+c)}$

- CI for TPF: $\left( \frac{\exp(\beta_L)}{1 + \exp(\beta_L)}, \frac{\exp(\beta_U)}{1 + \exp(\beta_U)} \right)$
Syphilis Test - adjusted TPF

\[ TPF_{adj} = 44.8\% \]

\[ \hat{\beta} = \log(0.448/0.552) = -0.209 \]

\[ \text{var}(\hat{\beta})= [770/ (21\times 749) + 8/(13\times 21) + 275/ (6\times 281)] = 0.491^2 \]

95% CI for \( \beta \): 
\[ -0.209 \pm 1.96 (0.491) = (-1.170, 0.756) \]

95% CI for \( TPF_{adj} \):
\[ (e^{-1.170}/[1+ e^{-1.170}], e^{0.756}/[1+ e^{0.756}]) = (0.237, 0.680) \]
Verification function of other factors

- verification may depend on other factors
  - Other tests
  - Family history
  - Previous test results

- For MAR to hold, must collect data & adjust for all factors \(A\) affecting verification

\[
TPF_{adj} = \frac{\sum_{A} P(D+ | Y+, A)P(Y+, A)}{\sum_{A} P(D+ | Y+, A)P(Y+, A) + P(D+ | Y-, A)P(Y-, A)}
\]
Non-ignorable verification bias

- Non-ignorable (NI) missing data mechanism – depends on unobserved factors related to disease
  - Physician decision based on perceived risk
  - Patient refusal, dropout
Accounting for NI verification bias

- \[ P(V,Y,D|A) = P(D|A) \ P(Y|D,A) \ P(V|Y,D,A) \]
- Let \( z_{mi} = i^{th} \) row of design matrix of model \( m \)
- Disease: logit \( P(D_i=1|A_i) = z_{0i}' \alpha \)
- Test: logit \( P(Y_i=1|D_i, A_i) = z_{1i}' \beta \)
- Missing data: logit \( P(V_i=1|D_i, Y_i, A_i) = z_{2i}' \gamma \)
- Test MAR by testing \( \gamma = 0 \) for \( D \)

Kosinski & Barnhart (2003)
Accounting for NI verification bias

- MLE obtained using EM algorithm
- Non-identifiability of parameters →
  - Limited range of models can be fit
- Sensitivity analysis – vary $\gamma$ for $D$ around estimated values & maximize remaining parameters

Kosinski & Barnhart (2003)
Selected Verification - Non-binary Tests

• Continuous audiology tests for neonatal hearing impairment
  – GS is a behavioral test at 1 year of age
    • expensive & difficult to obtain

• Electron beam computed tomography (EBCT) for myocardial ischemia
  – GS is single-photon emission CT (SPECT)
    • More expensive, requires ingestion of radioactive tracing material
Verification bias – continuous tests

- \( \uparrow Y \) more likely verified, apparent TPF\( \uparrow \) & FPF\( \uparrow \)
- Biased operating points on ROC
- curve
Two-phase design

- Phase 1: test result (Y) & additional information (A) on all study subjects
- Phase 2: disease status (D) for a subset (V=1) where P(V=1) may depend on phase 1 data
- Key observation – TPF(c) & FPF(c) are ratio of prevalence probabilities (Alonzo & Pepe, 2005)

\[
TPF(c) = \frac{P(Y \geq c, D=1)}{P(D=1)}, \quad FPF(c) = \frac{P(Y \geq c, D=0)}{P(D=0)}
\]
Steps

1. Obtain bias-corrected estimates of TPF(c) & FPF(c) for all c
2. Estimate ROC curve by plotting bias-corrected TPF(c) & FPF(c)
3. Estimate summary index
   a. Empirical estimate of AUC or pAUC
   b. Bootstrap to obtain CI
Imputation Methods

- Full imputation (FI); mean score imputation (MSI)

\[
\hat{T}_F = \frac{\sum_{i=1}^{n} I[Y_i \geq c] \hat{\rho}_i}{\sum_{i=1}^{n} \hat{\rho}_i}
\]

\[
\hat{T}_M = \frac{\sum_{i=1}^{n} I[Y_i \geq c] \{V_i D_i + (1-V_i)\hat{\rho}_i\}}{\sum_{i=1}^{n} \{V_i D_i + (1-V_i)\hat{\rho}_i\}}
\]

- \(\rho_i = P(D_i=1|Y,A)\) can be estimated using logistic regression applied to verification sample
- Inconsistent under model misspecification
- Extension of Begg and Greenes approach
Re-weighting Methods

• Inverse probability weighting (IPW) – extension of binary approach

• Semi-parametric efficient (SPE)

\[
\hat{TPF}_{IPW} = \frac{\sum_{i=1}^{n} I[Y_i \geq c] \frac{V_i D_i}{\hat{\pi}_i}}{\sum_{i=1}^{n} \frac{V_i D_i}{\hat{\pi}_i}}
\]

\[
\hat{TPF}_{SPE} = \frac{\sum_{i=1}^{n} I[Y_i \geq c] \left\{ \frac{V_i D_i}{\hat{\pi}_i} - (V_i - \hat{\pi}_i) \hat{\rho}_i / \hat{\pi}_i \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_i}{\hat{\pi}_i} - (V_i - \hat{\pi}_i) \hat{\rho}_i / \hat{\pi}_i \right\}}
\]

• \( \pi_i = P(V_i = 1|Y,A) \) may be known or may need to be estimated
• IPW inconsistent under model misspecification
• SPE consistent if either disease or verification models are correct
Neonatal Hearing Screening Impairment Study

- Screened for hearing loss using distortion product otoacoustic emissions (DPOAE)
- Hearing impairment defined by behavioral test
- 5101 ears (147 impaired) on 2762 infants
- Simulate verification status as function of DPOAE
  - Verify all subjects w/ at least one DPOAE >80th quantile of DPOAE values
  - Verify rest w/ probability 0.4
  - 43% fewer infants require GS

Norton et al. (2000)
Bias correction

(a)  
(b)  
(c)  
(d)
## AUC estimates

<table>
<thead>
<tr>
<th>Estimator</th>
<th>AUC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full data</td>
<td>0.632</td>
<td>(0.582, 0.683)</td>
</tr>
<tr>
<td>SPE</td>
<td>0.632</td>
<td>(0.581, 0.710)</td>
</tr>
<tr>
<td>IPW</td>
<td>0.632</td>
<td>(0.578, 0.709)</td>
</tr>
<tr>
<td>MSI</td>
<td>0.648</td>
<td>(0.593, 0.705)</td>
</tr>
<tr>
<td>Complete case</td>
<td>0.665</td>
<td>(0.603, 0.728)</td>
</tr>
</tbody>
</table>

Accounting for NI missingness

- Ordinal test results - Estimate ROC curve and AUC assuming binormal model [Zhou & Rodenberg, 1998]

- Continuous test results – estimate AUC using double robust estimator (similar to the SPE estimator) [Rotnitzky et al, 2006]
Paired Screen Positive (PSP) Design

• some settings only subjects + at least 1 of 2 tests, A & B, receive the GS test

• prostate cancer screening – PSA and DRE
  – only PSA + or DRE + receive biopsy
## PSP Data

<table>
<thead>
<tr>
<th></th>
<th>GS+</th>
<th>GS-</th>
</tr>
</thead>
<tbody>
<tr>
<td>B+</td>
<td>a</td>
<td>e</td>
</tr>
<tr>
<td>B-</td>
<td>b</td>
<td>f</td>
</tr>
<tr>
<td>A+</td>
<td>c</td>
<td>g</td>
</tr>
<tr>
<td>A-</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Paired Screen Positive Design

- Unable to estimate absolute accuracy
- Unbiased estimates of relative accuracy

\[
\begin{align*}
    r_{TPF} &= \frac{\Pr(A+ | D+)}{\Pr(B+ | D+)} = \frac{\Pr(A+ | D+) \Pr(D+)}{\Pr(B+ | D+) \Pr(D+)} = \frac{\Pr(A+, D+)}{\Pr(B+, D+)} = DR_A \\
    r_{FPF} &= \frac{\Pr(A+ | D-)}{\Pr(B+ | D-)} = \frac{\Pr(A+ | D-) \Pr(D-)}{\Pr(B+ | D-) \Pr(D-)} = \frac{\Pr(A+, D-)}{\Pr(B+, D-)} = FR_A
\end{align*}
\]
Relative Accuracy Estimators

<table>
<thead>
<tr>
<th></th>
<th>GS+</th>
<th></th>
<th>GS-</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

| A+   | | | |
|-----|---|---|
| a   | b | e | f |
| c   | ? | g | ? |

\[
r_{TPF} = \frac{(a+b)}{(a+c)}
\]

\[
r_{FPF} = \frac{(e+f)}{(e+g)}
\]

\[
\text{var}(\log r_{TPF}) = \frac{(b+c)}{(a+b)(a+c)}
\]

\[
\text{var}(\log r_{FPF}) = \frac{(f+g)}{(e+f)(e+g)}
\]
### Prostate Cancer in Black Men

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th></th>
<th>No Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DRE+</td>
<td>DRE-</td>
<td>DRE+</td>
<td>DRE-</td>
</tr>
<tr>
<td>PSA+</td>
<td>10</td>
<td>28</td>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>PSA-</td>
<td>8</td>
<td>?</td>
<td>26</td>
<td>?</td>
</tr>
</tbody>
</table>

\[
\hat{r}_{TPF}(PSA, DRE) = \frac{10 + 28}{10 + 8} = 2.1 \quad 95\% \text{ CI: 1.3, 3.3}
\]

\[
\hat{r}_{FPF}(PSA, DRE) = \frac{3 + 38}{3 + 26} = 1.4 \quad 95\% \text{ CI: 0.9, 2.2}
\]

Smith et al. (1997)
Selected Verification Studies

• Not by design
  • dropout or subject refusal
  • Different referral rates

• By design – 2 phase studies
  – Phase 1: relatively cheap test
  – Phase 2: expensive GS
  – In low prevalence studies, variance of TPF of most concern
Two-phase studies

- For TPF, 2-phase preferred to SRS only if ratio of costs $\geq 50$ & prev $\geq 0.5$ [McNamee, 2002]
- If all test+ in phase 1 get GS, then selected verification of test- only beneficial (modest savings) when all of these are met:
  - TPF $< 70$
  - FPF $> 20$
  - Ratio of costs $> 5$
  [Obuchowski & Zhou, 2002]
Summary

• Selected disease verification can result in biased accuracy estimates if not properly accounted for

• Relative accuracy (not absolute accuracy) can be estimated in screen positive studies

• Little efficiency gain in selected verification studies
References (1)


References (2)


Adjusting for Covariates in ROC Analysis
Outline

a. Motivation for covariate adjustment
b. Methods for covariate adjustment
c. Settings where covariate adjustment is important
d. Contrast covariate adjustment with other uses for covariates
e. ROC regression
Confounding Example: $Y$ varies with $Z$ (e.g., center), $D$ varies with $Z$ (center), and the unadjusted ROC is biased too high (scenario 1)

Scenario 1: $P[Z = 1|D = 0] = 0.10$ and $P[Z = 1|D = 1] = 0.50$
Bias Even When $D \perp Z$

- when $Z$ affects marker observations equally in cases and controls but is independent of $D$
  - e.g., in a multi-center study where the case-control ratio is held constant across centers
  - e.g., a normal or abnormal condition in controls may affect marker levels, but be unrelated to disease risk
- The unadjusted ROC is biased (attenuated). Result 6.1, Pepe book.
**Example:** $Y$ varies with $Z$ (e.g., center) $D$ is independent of $Z$ (center), and the unadjusted ROC curve is attenuated (scenario 2)

Scenario 2: $P[Z = 1|D = 0] = P[Z = 1|D = 1] = 0.50$
When Covariates Affect Discrimination

When $Z$ affects the separation between the $Y_D$ and $Y_{\bar{D}}$ distributions

- the ROC curve varies with $Z$
- e.g., specimen storage time affects discrimination
- analogous to effect modification in association studies
$Z$ affects discrimination, and marker observations among controls.
Z affects discrimination, but not marker observations among controls.
For now, we focus on covariates that don’t affect discrimination

- $Y_D$ may vary with $Z$
- $D$ may vary with $Z$
- discrimination does not (no effect modification)
- i.e., $Z$ affects case and control distributions equally
Methods for Covariate Adjustment
Covariate Adjustment in Studies of Association

- when studying the association between exposure and outcome
- with $Z$ a confounder but not effect modifier
- quantify exposure-outcome association using adjusted odds ratio
- odds ratio for population with fixed $Z$ value
- e.g., odds ratio for smoking and lung cancer among subjects of the same age
Covariate Adjustment in Biomarker Studies

- covariate-adjusted ROC ($\text{AROC}$)
- ROC curve for population with fixed covariate value
- common covariate-specific ROC curve
- analogous to the adjusted odds ratio

$\text{AROC}(t) = P[Y_D > S_{\overline{D}Z_D}^{-1}(t)]$

- plot of the TPF vs. FPF for rules that classify using covariate-specific thresholds chosen to hold covariate-specific FPFs constant across covariate values

$\text{AROC}(t) = P[pv(Y_D | Z_D) \geq 1 - t]$

- CDF of case observations standardized with respect to covariate-specific control distribution
Summary Measures for the $\text{AROC}$

- $\text{AROC}(t)$ at a fixed FPF = $t$
  - percent of cases detected when common covariate-specific FPF is $t$

- $\text{AROC}^{-1}(u)$ at a fixed TPF = $u$
  - common covariate-specific FPF when $u\%$ cases detected

- $\text{AAUC} = \int_{0}^{1} \text{AROC}(t) \, dt$
  - probability of correctly ordering randomly chosen case and control marker observation with the same covariate value

- $p\text{AAUC}(t_0) = \int_{0}^{t_0} \text{AROC}(t) \, dt$
  - $p\text{AAUC}(t_0)/t_0$ is probability of correctly ordering randomly chosen case and control marker observation with the same covariate value, given the control observation is above the $1 - t_0$ quantile of the covariate-specific control distribution
Estimating the $\hat{\text{AROC}}$ Nonparametrically with Discrete $Z$

$$
\hat{\text{AROC}}(t) = \frac{\sum_{i=1}^{n_D} I(Y_{Di} \geq \hat{S}_{DZ_i}^{-1}(t))}{n_D}
$$

$$
= \frac{\sum_{i=1}^{n_D} I(\hat{pv}(Y_{Di}|Z_{Di}) \geq 1 - t)}{n_D}
$$

- use empirical marker distribution for controls in each category of $Z$ to generate thresholds or percentile values

- $\hat{\text{AROC}}$ is a weighted average of covariate-specific empirical ROC curves
Estimating the AROC For General $Z$

\[ \widehat{\text{AROC}}(t) = \frac{1}{n_D} \sum_{i=1}^{n_D} I[\widehat{pv}(Y_{D_i}|Z_{D1}) \geq 1 - t]/n_D \]

- for discrete, continuous, multiple $Z$s
- where a regression model fit to control data $(Y_{\bar{D}}, Z_{\bar{D}})$ yields $\widehat{pv}(y|Z)$ or covariate specific quantiles
• fit a parametric model to \((Y_{\bar{D}}, Z_{\bar{D}})\), e.g.,

\[
Y_{\bar{D}} = \mu(Z_{\bar{D}}) + \sigma(Z_{\bar{D}})\epsilon, \quad \epsilon \sim N(0, 1)
\]

• fit a semi-parametric model (Heagerty and Pepe 1999) to \((Y_{\bar{D}}, Z_{\bar{D}})\):

\[
Y_{\bar{D}} = \mu(Z_{\bar{D}}) + \sigma(Z_{\bar{D}})\epsilon
\]

where \(\mu(Z_{\bar{D}}) = E(Y_{\bar{D}}|Z_{\bar{D}})\) and \(\sigma(Z_{\bar{D}}) = \text{Var}(Y_{\bar{D}}|Z_{\bar{D}})\) are parametric but \(\epsilon \sim F_0\) is estimated non-parametrically as the empirical distribution of the residuals \(\frac{Y_{\bar{D}j} - \hat{\mu}(Z_{\bar{D}j})}{\hat{\sigma}(Z_{\bar{D}j})}\)

• \(pv(Y|Z) = \hat{F}_0((Y - \hat{\mu}(Z))/\hat{\sigma}(Z))\) or

\[
\hat{S}_{D,Z}^{-1}(t) = \hat{\mu}(Z) + \hat{\sigma}(Z)\hat{q}(1 - t)
\]
Estimating the AROC Parametrically

\[ \hat{AROC}(t) = \hat{P}[\hat{p}v(Y_{D_i} | Z_{D_i}) \geq 1 - t] \]

- \( \hat{P} \) based on parametric form for the AROC, e.g., a binormal model
  \[ AROC(t) = \Phi(a + b\Phi^{-1}(t)) \]

- results in a smooth estimated ROC curve
Stata Software

`roccurve` for estimating and plotting AROC curves

`comproc` for comparing AROC curves
Choices Made in AROC Estimation

1. Covariates to be used for adjustment (adjcov)

2. Modeling covariate effects on $Y_D$ (adjmodel)
   - stratify (for discrete $Z$) (adjm(strat))
   - assume $Z$ acts linearly on $Y_D$ (adjm(linear))

3. Calculating percentile values (pvcmethod)
   - empirical estimate of $Y_{DZ}$ distribution (pvc(empirical))
   - assume normal $Y_{DZ}$ distribution (pvc(normal))

4. Calculating ROC curve (rocmethod)
   - estimate empirically (rocmethod(empirical))
   - assume binormal ROC (rocmethod(binormal))
Summary Measures Reported

- $\hat{AROC}(t_0), \hat{AAUC}, p\hat{AAUC}(t_0)$
- standard errors based on bootstrapping
  - options for case-control, within covariate strata, or clustered sampling
- comparisons between markers based on Wald tests, e.g.,

$$\frac{\hat{AROC}_1(t_o) - \hat{AROC}_2(t_o)}{se(\hat{AROC}_1(t_o) - \hat{AROC}_2(t_o))}$$
Example: Audiology Data

- study of hearing impairment in newborns
- each of 3 audiology tests given to 2742 babies
- gold standard: behavioral test at 9-12 months
- age, gender may affect test results among controls
`roccurve d y1, adjcov(male) cl(id) roct(0.2) ci`

- calculates $\mathcal{A}$ROC for $Y_1$, adjusted for gender
- by stratifying (default)
- estimates control distribution empirically (default)
- estimates ROC curve empirically (default)
- weighted average of gender-specific empirical ROC curves
- bootstrap for standard errors, clustered on ID
- calculates confidence interval for $\mathcal{A}$ROC (0.2)

marker: DPOAE 65 at 2kHz
‘roccurve d y1, adjcov(currage male) adjm(linear) pvc(normal)
rocm(binormal) cl(id) roct(0.2) ci’

- calculates AROC for $Y_1$, adjusted for age and gender
- by fitting a linear regression model for $Y_1$ among controls as a function of age and gender
- assumes normal $Y_1$ control distribution
- estimates binormal ROC curve
Settings Where Covariate Adjustment is Important
Covariate adjustment is important whenever

- $Y_D$ varies with $Z$
- unadjusted ROC is biased
Covariate adjustment important when comparing markers

- ROC curves can be differentially biased, leading to faulty marker comparisons
**Example:** $Y_1$ and $Y_2$ have the same performance, $Y_1$ varies with $Z$, but $Y_2$ does not. The unadjusted ROC curve for $Y_1$ is attenuated while that for $Y_2$ is not.
Contrasting Covariate Adjustment With Other Uses for Covariates
Risk Score Estimation is Different From Covariate Adjustment

• risk score: \( R \equiv P[D = 1|Y, Z] \)

• commonly estimated using logistic regression:

\[
\log \text{odds} P[D = 1|Y, Z] = \beta_0 + \beta_1 Y + \beta_2 Z
\]

• the ROC curve for \( R \) describes the ability of \( Y \) and \( Z \) together to discriminate
  
  – \( Z \) allowed to contribute to discrimination
  
  – may perform well even if \( Y \) is a poor classifier, if \( Z \) is a good classifier

• the \( AROC \) describes the ability of \( Y \) to discriminate in a population with fixed \( Z \)
(a) adjusted ROC

(b) combination marker

CA 125 | CA 19–9

CA 125 + CA 19–9
Incremental Value Estimation is Different From Covariate Adjustment

- incremental value: the amount of discriminatory accuracy of $Y$ over and above $Z$

- compare the ROC curve for $R$ to the ROC curve for $Z$ alone
  - $R$ is the optimal combination of $Y$ and $Z$ for discrimination (McIntosh and Pepe 2002)

- the AROC does not represent the incremental value of $Y$
  - very specific combination of $Y$ and $Z$, in general unrelated to the risk score
Roles for Covariates

• contribute to discrimination: model the risk score, $P[D = 1|Y,Z]$
  – e.g., other markers

• baseline predictors for evaluating incremental value of $Y$: compare ROC for $(Y,Z)$ to ROC for $Z$ alone
  – e.g., risk factor information or other markers

• affect marker distributions but not discrimination: $\Delta$ROC
  – e.g., study site

• affect discrimination: evaluate covariate-specific ROC curves
  – e.g., disease severity
ROC Regression

- continuous marker, $Y$

- $Z$ affects the separation between the $Y_D$ and $Y_{\bar{D}}$ distributions (discrimination)
  - e.g., $Z =$ expertise of test operator, test location, specimen storage time
  - e.g., disease severity

- $Z$ may or may not affect marker values among controls
$Z$ affects discrimination, but not marker observations among controls.
$Z$ affects discrimination and marker observations among controls
When $Z$ affects discrimination

- identify covariate effects to optimize test performance

\[ \text{AROC}(t) = \int \text{ROC}_Z(t) dh_D(Z), \] still interpretable
  
  - weighted average of covariate-specific ROC curves
  
  - analogous to Mantel-Haentzel odds ratio

- but estimates of covariate-specific ROC curves are of interest
Binormal ROC-GLM Model

\[ \text{ROC}_Z(t) = \Phi(\alpha_0 + \alpha_1 \Phi^{-1}(t) + \alpha_2 Z) \]

- stipulates that the ROC curve is binormal for each fixed \( Z = z \) with slope parameter \( \alpha_1 \) and intercept \( \alpha_0 + \alpha_2 z \)
- if \( \alpha_2 > 0 \), the ROC curve is higher for larger \( Z \)
- if \( \alpha_2 = 0 \), this reduces to a binormal AROC model (covariates are only used for adjustment)
Example: Audiology Data

- study of hearing impairment in newborns
- each of 3 audiology tests given to 2742 babies
- gold standard: behavioral test at 9-12 months
- age, gender may affect test results among controls
- assess whether the ROC curve depends on age
ROC-GLM Model

\[
Y_D = \beta_0 + \beta_1 Z_{male} + \beta_2 Z_{age} + \epsilon, \quad \epsilon \sim N(0, 1)
\]

\[
\text{ROC}_Z(t) = \Phi(\alpha_0 + \alpha_1 \Phi^{-1}(t) + \alpha_2 Z_{age})
\]

- marker observations are allowed to depend on age and gender
- the ROC curve is allowed to depend on age
  - binormal ROC at each age with intercept = \(\alpha_0 + \alpha_2 \text{ age}\)
    and slope = \(\alpha_1\)
Fitting the Model

• choose FPFs $t_1, \ldots t_K$ at which to estimate the ROC curve
• calculate the corresponding thresholds $y_k = S_{D,Z}^{-1}(t_k)$
• for each case observation, create binary indicator variables $U_k = I[Y_D > y_k]$
• perform probit regression with outcome $U_k$ and covariates $\Phi^{-1}(t_k)$ and $Z_{age}$
  - fits the model

$$ROC_Z(t) = P[Y_D > y_k] = \Phi(\alpha_0 + \alpha_1 \Phi^{-1}(t) + \alpha_2 Z_{age})$$
Stata Software

- *rocreg*

- fits binormal ROC-GLM models
Choices Made in Binormal ROC-GLM Estimation

1. Covariates to be used for adjustment (adjcov)

2. Modeling covariate effects on $Y_D$ (adjmodel)
   - stratify (for discrete $Z$) (adjm(strat))
   - assume $Z$ acts linearly on $Y_D$ (adjm(linear))

3. Calculating placement values, $S_{DZ}(YD)$ (pvcmethod)
   - empirical estimate of $Y_{DZ}$ distribution (pvc(empirical))
   - assume normal $Y_{DZ}$ distribution (pvc(normal))

4. Covariates to include in the ROC model (regcov)
‘rocreg d y1, regcov(currage) adjcov(currage male) adjm(linear) cl(id)’

- fits binormal ROC-GLM for $Y_1$
- marker values among controls adjusted for age, gender
- by fitting a linear regression model for $Y_1$ among controls as a function of age and gender
- estimates control distribution empirically (default)
- allows the ROC curve to depend on age
- bootstrap for standard errors, clustered on ID
**ROC Regression Results**

\[ \text{ROC} - Z(t) = (\alpha_0 + \alpha_1 \Phi^{-1}(t) + \alpha_2 Z_{age}) \]

<table>
<thead>
<tr>
<th>Factor</th>
<th>Parameter</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>( \alpha_2 )</td>
<td>0.05</td>
<td>(-0.01, 0.10)</td>
</tr>
<tr>
<td>ROC Intercept</td>
<td>( \alpha_0 )</td>
<td>-1.27</td>
<td>(-3.52, 0.96)</td>
</tr>
<tr>
<td>ROC slope</td>
<td>( \alpha_1 )</td>
<td>0.94</td>
<td>(0.78, 1.09)</td>
</tr>
</tbody>
</table>

- positive age coefficient \( \Rightarrow \) accuracy increases with age
- not statistically significant
Summary

- adjust for covariates associated with control marker observations
  - even when comparing markers
  - even in matched studies

- AROC, the common covariate-specific ROC curve

- covariate adjustment is different from:
  - ROC for adjusted risk score
  - assessment of incremental value

- ROC regression to evaluate covariate effects on discrimination
Bibliography


Study Design Part II
Setting

- continuous marker, \( Y \)
- case-control study
- covariate (\( Z \)) affects marker observations and is associated with the outcome
  - eg age impacts PSA levels, and is a risk factor for prostate cancer
- to begin, assume \( Z \) does not affect discrimination (the ROC curve)
- should the cases and controls be matched with respect to \( Z \)?
Matched Case-Control Design

• randomly sample cases

• sample controls matched to cases with respect to $Z$
  – ensures cases and controls have the same distribution of $Z$

• attempts to eliminate the contribution of covariates to discrimination (confounding)
Example: The Physicians’ Health Study

- randomized, placebo-controlled study of aspirin and β-carotene among 22,071 men
- sub-study assessing accuracy of PSA as prostate cancer screening tool
- serum sample taken at enrollment, stored
- 429 prostate cancer cases (≤ 12 years after enrollment)
- 1,287 age-matched controls (for each case, 3 within 1 year of age)
- matching an attempt to eliminate the contribution of age: cases tend to be older, older subjects higher PSA
  - without adjustment for age, ROC for PSA overly optimistic
- what are the implications of the matched design?
Outline

• review matching in association studies

• matching in biomarker studies:
  – covariate-adjusted analysis is essential
  – precludes assessment of incremental value
  – often an efficient design

• sample size calculations for matched biomarker studies
Matching in Case-Control Studies of Association

- study the association between an exposure and disease

- matching necessitates adjusting for the matching covariates in the analysis
  - attenuates crude unadjusted odds ratio

- estimate adjusted odds ratio: the odds ratio among population with fixed covariate value
Example: Distribution of cases and person-years in cohort study

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th></th>
<th>Females</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
<td>Non-exposed</td>
<td>Exposed</td>
<td>Non-exposed</td>
</tr>
<tr>
<td>Diseased</td>
<td>450</td>
<td>10</td>
<td>25</td>
<td>45</td>
</tr>
<tr>
<td>Person-years</td>
<td>90,000</td>
<td>10,000</td>
<td>10,000</td>
<td>90,000</td>
</tr>
<tr>
<td>Rate ($\times 10^{-3}$)</td>
<td>5.0</td>
<td>1.0</td>
<td>2.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Rate ratio</td>
<td>5.0</td>
<td></td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>Crude rate ratio</td>
<td>($\frac{450+25}{100,000}/\frac{10+45}{100,000}$) = 8.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- gender confounds the exposure-disease association

Modified from Breslow (2005) *Handbook of Epidemiology*
## Expected distributions in matched case-control study

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
<td>Non-exposed</td>
</tr>
<tr>
<td>Cases</td>
<td>450</td>
<td>10</td>
</tr>
<tr>
<td>Controls</td>
<td>414</td>
<td>46</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>5.0</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{crude odds ratio} = \frac{(450+25)\times(46+63)}{(10+45)\times(414+7)} = 2.2
\]

- the crude odds ratio in the gender-matched study overcompensates
- the gender-specific odds ratios correctly estimate the common rate ratio

Modified from Breslow (2005) *Handbook of Epidemiology*
Matching in Case-Control Studies of Association

• matching necessitates adjusting for the matching covariates in the analysis

• matching increases the efficiency of the adjusted odds ratio estimator

• precludes estimation of the association between matching covariates and disease
Analysis of Matched Biomarker Studies

• unadjusted analysis: ignores the matching
• adjusted analysis: adjust for the matching covariates
• adjusted analysis is essential
Unadjusted Analysis

• the unadjusted ROC curve: the traditional pooled ROC curve in the matched data

• pools all cases together, regardless of $Z$, all controls together, regardless of $Z$

• no real-world value—artificial control set
  – FPF doesn’t relate to any real population of controls

• analogous to the crude unadjusted OR in matched association study
Adjusted Analysis

- adjust for the matching covariates
- the covariate-adjusted ROC curve (AROC)
- real-world value: the performance of the marker in a population with fixed covariate value
- analogous to the adjusted OR in matched association study
Example

\[ Z = 0 \]

\[ Z = 1 \]

density

unmatched data

matched data

ROC curves

- AROC
- pooled unmatched
- pooled matched

TPF

FPF
Result: the unadjusted ROC is attenuated relative to the AROC

- implies that matching does not in and of itself control for confounding
- matching necessitates covariate adjustment: unadjusted ROC is biased
- directly analogous to result in association setting
The Physicians’ Health Study

- empirical unadjusted ROC curve and binormal AROC curve
An adjusted analysis is also essential when comparing markers

- the unadjusted ROC curves can be differentially biased, leading to faulty marker comparisons
**Example:** $Y_1$ and $Y_2$ have the same performance, $Y_1$ varies with $Z$, but $Y_2$ does not. The unadjusted ROC curve for $Y_1$ is attenuated while that for $Y_2$ is not.
When the Matching Covariate Affects Discrimination

- eg $Z = \text{specimen storage time}$

- $A\text{ROC}$ still interpretable
  
  $A\text{ROC}(t) = \int \text{ROC}_Z(t) \ h_{ZD}$

- ultimately want to estimate covariate-specific ROC curves ($\text{ROC regression}$)

- the unadjusted ROC in a matched study is difficult to interpret and not directly relevant
**Key Point:** In a matched study, the analysis must adjust for the matching covariates
Matching and Incremental Value

- incremental value: the amount of discriminatory accuracy of $Y$ over and above $Z$

- compare the ROC curve for $R = P[D = 1|Y, Z]$ to the ROC curve for $Z$ alone
  
  - $R$ is the optimal combination of $Y$ and $Z$ for discrimination (McIntosh and Pepe 2002)

- eg, compare the ROC curve for age alone to the ROC curve for the risk score

  \[
  \text{log odds } P(D = 1|\text{age, PSA}) = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{PSA}
  \]
Matching Precludes Direct Evaluation of Incremental Value

• matching $\implies Z$ has the same distribution in controls and cases
• $Z$ is useless as a classifier in the matched data
• cannot directly assess incremental value of $Y$ over $Z$
• neither does the performance of $Y$ in the matched data reflects its incremental value
Examples: (a) The incremental value is large, and the $A$ROC and unadjusted ROC curves are low. (b) The incremental value is small, and the $A$ROC and unadjusted ROC curves are high.
Matching and Efficiency

Consider broad class of designs: cases randomly sampled, controls sampled with respect to covariate-dependent sampling distribution

- eg random case control sampling (ignore $Z$ when sampling controls)
- eg matching (sample controls to ensure the same distribution of $Z$ in cases and controls)
Result: matching is the most efficient control sampling strategy

- minimizes asymptotic $\text{Var}(\hat{A\text{ROC}}(t))$
- when $Z$ does not affect discrimination
  - when $Z$ affects discrimination, optimal design is more complex
- optimal matching ratio

$$\frac{n_D}{n_{\bar{D}}} = \frac{1}{\frac{\partial}{\partial t}A\text{ROC}(t)} \cdot \sqrt{\frac{A\text{ROC}(t)(1 - A\text{ROC}(t))}{t(1 - t)}}$$

- matching cannot lead to loss of efficiency
  - even if $Y$ and $Z$ are independent
  - in contrast to association studies
Example: Physicians’ Health Study

- age does not affect the discriminatory accuracy of PSA
- matching on age is the optimal design
- what should the matching ratio be?
Sample Size Calculations for Matched Case-Control Studies
Matched Case-Control Design

- design based on covariate-adjusted analysis ($\mathcal{AROC}$)
- assuming discrete covariate $Z = 1, \ldots, K$
- choose minimally acceptable $(\text{FPF}_0, \text{TPF}_0)$
- does the TPF of the test exceed $\text{TPF}_0$ at the threshold corresponding to $\text{FPF}_0$?
  - ie is $\mathcal{AROC}(\text{FPF}_0) \geq \text{TPF}_0$
- test is minimally accurate if $\mathcal{AROC}(\text{FPF}_0) \geq \text{TPF}_0$ with high confidence
  - if lower limit of confidence interval for $\mathcal{AROC}(\text{FPF}_0)$ exceeds $\text{TPF}_0$
Sample Size Formula Based on Asymptotic Variance

For testing

\[ H_0 : \mathcal{AROC}(\text{FPF}_0) \leq \text{TPF}_0 \]

in order to achieve type-I error \( \alpha \) and power \( 1 - \beta \) at \( \text{TPF}_1 = \mathcal{AROC}(\text{FPF}_0) \) we require

\[
n_D = \frac{\left\{ \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta) \right\}^2 \ V_1}{(\text{TPF}_0 - \text{TPF}_1)^2}
\]

where \( V_1 = n_D \ \text{Var}(\hat{\mathcal{AROC}}_e(\text{FPF}_0)) \)

- based on lower limit of one-sided confidence interval for \( \mathcal{AROC}(\text{FPF}_0) \)
\[ V_1 = n_D \text{ Var}(\widehat{\text{AROC}}) \]

\[ = \text{AROC}(t)(1 - \text{AROC}(t)) + \lambda \left\{ \frac{\partial \text{AROC}(t)}{\partial t} \right\}^2 t(1 - t) \]

- \(\lambda = n_D/n_D\), the case-control ratio constant across \(Z\) (matching ratio)

- \(\lambda_{opt} = \left\{ \frac{\partial \text{AROC}(t)}{\partial t} \right\}^{-1} \sqrt{\frac{\text{AROC}(t)(1 - \text{AROC}(t))}{t(1 - t)}}\) (Janes and Pepe, 2007)

Input:

- \(t = \text{FPF}_0\)

- \(\text{AROC}(t) = \text{TPF}_1\)

- \(\frac{\partial \text{AROC}(t)}{\partial t}\) at \(t = \text{FPF}_0\)
Example: Physicians’ Health Study, cont’d

- design future study to evaluate discriminatory accuracy of PSA
- using PHS as pilot data
- $FPF_0 = 0.025$, $TPF_0 = 0.15$ minimally useful
- $TPF_1 = 0.18$ from pilot data
- \[ \frac{\partial AROC(t)}{\partial t} = 4.36 \] from pilot data
- 90% power, 5% type 1 error
- $\lambda_{opt} = 0.57$, case-control ratio constant across age
- require $(n_D, n_{\overline{D}}) = (3515, 6206)$
Summary

- matching should be carefully considered in relation to questions of interest
- matching necessitates covariate adjustment
- matching often maximally efficient
  - optimal matching ratio depends on shape of AROC
- matching precludes assessment of incremental value of marker over matching covariates
- design of matched study:
  - characterize minimally useful test (null hypothesis)
  - identify desired performance (alternative hypothesis)
  - calculate matching ratio
  - calculate starting points for sample sizes using asymptotic formula
  - validate using simulations
References

Imperfect/Missing Reference Test

Todd A. Alonzo, PhD
University of Southern California
talonzo@childrensoncologygroup.org

July 29, 2007
Outline

1. Effect of imperfect reference test on TPR and FPR

2. Latent Class Analysis

3. Bayesian Approach

4. Composite reference standard
Setting

• Ideally all study subjects receive screening test and gold standard (GS) - provides **definitive** determination of D

• Often gold standard not perfect – reference standard/bronze standard
  – e.g., Cell culture used to be GS for chlamydia, malaria, gonorrhea infections
  – e.g., cancer – biopsy may not be from correct location
  – e.g., questionnaire for alcoholism – alcoholics may not respond truthfully
Setting

- Tests being developed that more accurate than “GS”
- Many so-called FP are TP
- Imperfect GS results in biased accuracy estimates
Reference test (R) misses disease 20% of time

<table>
<thead>
<tr>
<th>Truth</th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y+</td>
<td>80</td>
<td>30</td>
</tr>
<tr>
<td>Y-</td>
<td>20</td>
<td>70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observe</th>
<th>R+</th>
<th>R-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>74</td>
</tr>
</tbody>
</table>

prev = \( \frac{100}{200} = 0.5 \)  
TPF = \( \frac{80}{100} = 0.8 \)  
FPF = \( \frac{30}{100} = 0.3 \)

prev = \( \frac{80}{200} = 0.4 \)  
TPF = \( \frac{64}{80} = 0.8 \)  
FPF = \( \frac{46}{120} = 0.38 \)
Reference test (R) misses disease 20% of time - in truly diseased when Y-

<table>
<thead>
<tr>
<th>Truth</th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y+</td>
<td>80</td>
<td>30</td>
</tr>
<tr>
<td>Y-</td>
<td>20</td>
<td>70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observe</th>
<th>R+</th>
<th>R-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>90</td>
</tr>
</tbody>
</table>

prev = 100/200 = 0.5
TPF = 80/100 = 0.8
FPF = 30/100 = 0.3

prev = 80/200 = 0.4
TPF = 80/80 = 1.0
FPF = 30/120 = 0.25
New test has perfect accuracy. Reference test has TPR = 0.9, FPR=0.1

<table>
<thead>
<tr>
<th>Truth</th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y+</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Y-</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observe</th>
<th>R+</th>
<th>R-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

prev = \(\frac{100}{200} = 0.5\)

TPF = \(\frac{100}{100} = 1.0\)

FPF = \(\frac{0}{100} = 0\)

prev = \(\frac{100}{200} = 0.5\)

TPF = \(\frac{90}{100} = 0.9\)

FPF = \(\frac{10}{100} = 0.1\)
Latent Class Analysis

- Unobserved latent binary variable $D$ indicates presence ($D=1$) or absence ($D=0$) of condition
- Observe $K$ binary tests $\{Y_1, ..., Y_K\}$
- Reference test may be one of these
- Model $P_\theta(Y_1, ..., Y_K | D)$
- Data: $Y_i = \{Y_{1i}, Y_{2i}, ..., Y_{Ki}\}$, $i=1, ..., n$ subjects
- $\theta$ can be estimated by maximizing likelihood if model has sufficient structure
Conditional Independence

- conditional independence
  - given $D$, $\{Y_1, \ldots, Y_K\}$ are statistically independent

$$P(Y_{1i}, \ldots, Y_{Ki} \mid D) = \prod_{k=1}^{K} P(Y_{ki} \mid D_i)$$

- $2^K - 1$ Degrees of freedom (dfs)
- $2K+1$ parameters: prevalence, TPF & FPF for each test
- Identifiable model for $K \geq 3$
Likelihood

\[
= \prod_{i=1}^{n} P(Y_i)
\]

\[
= \prod_{i=1}^{n} \left\{ P(Y_i | D_i = 1) \right. P(D_i = 1) + P(Y_i | D_i = 0) P(D_i = 0) \right\}
\]

\[
= \prod_{i=1}^{n} \left\{ \rho \prod_{k=1}^{3} P(Y_{ik} | D_i = 1) + (1 - \rho) \prod_{k=1}^{3} P(Y_{ik} | D_i = 0) \right\}
\]

\[
= \prod_{i=1}^{n} \left\{ \rho \prod_{k=1}^{3} \phi_{ik}^{Y_{ik}} (1 - \phi_k)^{1-Y_{ik}} + (1 - \rho) \prod_{k=1}^{3} \psi_{ik}^{Y_{ik}} (1 - \psi_k^{1-Y_{ik}}) \right\}
\]

\[
\phi_k = TPF_k \quad \quad \psi_k = FPF_k \quad \quad \rho = P(D = 1)
\]

\[
K = 3 \quad \Rightarrow \quad \text{7 parameters}
\]
## Hearing impairment test results

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>207</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>75</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>85</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>85</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>162</td>
</tr>
</tbody>
</table>

Pepe (2003)
Hearing testing – results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>LCA</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence</td>
<td>.54</td>
<td>.42</td>
</tr>
<tr>
<td>TPF A</td>
<td>.84</td>
<td>.66</td>
</tr>
<tr>
<td>TPF B</td>
<td>.76</td>
<td>.62</td>
</tr>
<tr>
<td>TPF C</td>
<td>.90</td>
<td>.75</td>
</tr>
<tr>
<td>FPF A</td>
<td>.13</td>
<td>.40</td>
</tr>
<tr>
<td>FPF B</td>
<td>.13</td>
<td>.36</td>
</tr>
<tr>
<td>FPF C</td>
<td>.31</td>
<td>.54</td>
</tr>
</tbody>
</table>
Conditional Dependence

• Parameter estimates biased if assume conditional independence & there is dependence [Vacek, 1985; Torrance-Rynard and Walter 1997]
Relaxing conditional independence

• Need $K \geq 4$ to include dependence & estimate parameters of interest

• Approaches to relax conditional independence
  – Log linear modeling [Epseland & Handelman, 1989]
  – Marginal approach for pairwise dependence [Yang & Becker, 1997]
  – Gaussian random effect [Qu, Tan, Kutner 1996]
  – Finite mixture model [Albert et al, 2001]
Criticisms of LCA

• Does not require formal disease definition
  – prev & test accuracy parameters not well-defined.

• Assumed latent class model not fully testable with observed data.

• Estimators biased for misspecified dependence structure [Albert & Dodd, 2004]

• Difficult to choose correct dependence structure [Albert & Dodd, 2004]
LCA insights

• Pepe & Janes (2007) - 3 tests and assume cond. indep.
• Provide closed form expressions for maximum likelihood parameter estimates
• TPF & FPF estimates start at \((Y=1) = TPF = FPF\) (uninformative test) estimate
  – Under cond indep, any correlation b/t tests due to common association with \(D\)
  – marginal + associations b/t pairs of tests ↑ estimated TPF from starting point
  – + associations b/t tests in observed data ↓ estimates of FPF from starting point
# Hearing impairment tests

\[ P(D=1|A, B, C) \]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>#</th>
<th>LCA</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>207</td>
<td>.9921</td>
<td>.6329</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>27</td>
<td>.8658</td>
<td>.5185</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>50</td>
<td>.7771</td>
<td>.4400</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>29</td>
<td>.1525</td>
<td>.2069</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>75</td>
<td>.8568</td>
<td>.3733</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>31</td>
<td>.2360</td>
<td>.3548</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>85</td>
<td>.1430</td>
<td>.3176</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>162</td>
<td>.0085</td>
<td>.2346</td>
</tr>
</tbody>
</table>

+ dependence inflates LCA estimates for + tests, deflates for - tests
Bayesian Approach

• Used when # parameters > # df’s [Joseph et al, 1995]
• Prior information used to distinguish b/t possible solutions for non-identifiable problem
• Data via likelihood function combined with prior distribution to derive posterior distributions
• Posterior distributions strongly dependent on prior information even with large sample sizes
• Extensions to account for conditional dependence [Dendukuri and Joseph, 2001]
  – Require conditional model specified completely
Composite Reference Standard

- several imperfect reference tests combined to form imperfect reference standard
- CRS should not depend on new test
- Examples:
  - CRS + if both ref. tests +
  - CRS + if either ref. test +

Alonzo & Pepe (1999)
**Chlamydia trachomatis testing of 3639 asymptomatic men**

Polymerase chain reaction (PCR) – DNA amplification

Ligase chain reaction (LCR) – DNA amplification

<table>
<thead>
<tr>
<th>Cell culture</th>
<th>PCR</th>
<th>Culture</th>
<th>LCR</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>225</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td></td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td></td>
<td>+</td>
<td>93</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td></td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td></td>
<td>+</td>
<td>24</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td></td>
<td>-</td>
<td>19</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
<td>+</td>
<td>34</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>3190</td>
</tr>
</tbody>
</table>

Johnson et al. (2000)
## Chlamydia “Either +” CRS

<table>
<thead>
<tr>
<th>PCR</th>
<th>C</th>
<th>LCR</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>225</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>93</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>24</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>19</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>34</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3190</td>
</tr>
</tbody>
</table>
Chlamydia “Both +” CRS

<table>
<thead>
<tr>
<th>PCR</th>
<th>C</th>
<th>LCR</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>225</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>93</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>24</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>19</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>34</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3190</td>
</tr>
</tbody>
</table>
## PCR Results

<table>
<thead>
<tr>
<th>Standard</th>
<th>Prev</th>
<th>TPF</th>
<th>FPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culture</td>
<td>8.1%</td>
<td>85.4%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Either +</td>
<td>11.6%</td>
<td>81.8%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Both +</td>
<td>6.8%</td>
<td>90.4%</td>
<td>4.3%</td>
</tr>
</tbody>
</table>
CRS

Advantages:
- several sources of information generate reference standard
- Reference standard indep of test investigated

Disadvantages:
- Still not perfect
- Requires 3rd test for test – on 2 other tests
  - Large # when prevalence is low
  - Modified design where only fraction re-tested
Summary

• Imperfect reference test can bias accuracy estimates

• Latent Class Analysis
  – Assumed latent class model not fully testable with observed data
  – Biased estimators when dependence structure is misspecified
  – Bayesian

• Composite reference standard
  – Not perfect
References


References


Combining Predictors and Incremental Value
Why Combine?

- multiple tests available;
  a combination *may* have improved performance over any one on its own

- existing markers/clinical predictors;
  what is increment in performance by having new test in addition to existing predictors?
How to Combine?

- Binary tests \((Y_A, Y_B)\)

<table>
<thead>
<tr>
<th>(Y_{CPH} = 0)</th>
<th>(Y_{CPH} = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_{EST} = 0)</td>
<td>151</td>
</tr>
<tr>
<td>(Y_{EST} = 1)</td>
<td>46</td>
</tr>
</tbody>
</table>

| \(Y_{EST} = 0\) | 25               | 183              |
| \(Y_{EST} = 1\) | 29               | 786              |
• the $OR$ rule: positive if ($Y_A = 1$ or $Y_B = 1$)

• “believe the positive”

• increases TPF and increases FPF, but by no more than $FPF_1 + FPF_2$

• used to combine specific non-sensitive tests

• useful if tests detect different diseased subjects
Example (CASS Data)

<table>
<thead>
<tr>
<th></th>
<th>$Y_{EST} = 0$</th>
<th>$Y_{EST} = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_{CPH} = 0$</td>
<td>151</td>
<td>46</td>
</tr>
<tr>
<td>$Y_{CPH} = 1$</td>
<td>176</td>
<td>69</td>
</tr>
<tr>
<td>$D = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_{EST} = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_{EST} = 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D = 1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_{EST} = 0$</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>$Y_{EST} = 1$</td>
<td>183</td>
<td>786</td>
</tr>
</tbody>
</table>

- The OR combination has TPF = $998/1023 = 97\%$
  
  FPF = $291/442 = 66\%$.

- Compare with $TPF_{EST} = 80\%$  
  $FPF_{EST} = 26\%$
• the $AND$ rule: positive if $(Y_A = 1$ and $Y_B = 1)$
• “believe the negative”
• decreases FPF and decreases TPF but to no less than $TPF_1 + TPF_2 - 1$
• used to combine sensitive non-specific tests
• useful if disease is present when all markers are positive
Example (CASS Data)

<table>
<thead>
<tr>
<th></th>
<th>$Y_{EST} = 0$</th>
<th>$Y_{EST} = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_{CPH} = 0$</td>
<td>151</td>
<td>46</td>
</tr>
<tr>
<td>$Y_{CPH} = 1$</td>
<td>176</td>
<td>69</td>
</tr>
<tr>
<td>$Y_{EST} = 0$</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>$Y_{EST} = 1$</td>
<td>183</td>
<td>786</td>
</tr>
</tbody>
</table>

- The AND combination has TPF = $786/1023 = 76.8\%$
- FPF = $69/442 = 15.6\%$
- Compare with TPF$_{EST} = 80\%$ FPF$_{EST} = 26\%$
More General Settings

- > 2 binary tests, the problem quickly becomes complex
- continuous tests
Principle

- the best performing combination tests are of the form:

  \[
  risk(Y) = \text{Prob}(D = 1|Y_A, Y_B, Y_C, \ldots) > \text{threshold}
  \]

- choose threshold so that in controls fraction with
  \( risk(Y) > \text{threshold} = f_0 \), some pre-specified FPF level

- then this rule has the highest TPF among all rules based on
  \((Y_A, Y_B, Y_C, \ldots)\) that have FPF= \( f_0 \)

- follows from Neyman-Pearson lemma (1933), which does not
  have an intuitive explanation (to me)
Implication

- using data on \( \{D, Y_A, Y_B, Y_C, \ldots \} \) for \( n \) subjects fit a model for \( P(D|Y_A, Y_B, \ldots) \)
- a “risk calculator” based on \( Y_A, Y_B, \ldots \)
- a monotone function of \( P(D|Y_A, Y_B, \ldots) \) works too
Logistic Regression Model

\[ \log\{P[D = 1|Y]/(1 - P[D = 1|Y])\} = \alpha_0 + \alpha_A Y_A + \alpha_B Y_B + \ldots \]

- the model can include transformations, interactions, and so is flexible
- available in all standard statistical packages
- well developed techniques for assessing model fit, comparing different models
- can be fit to case-control data since this only affects the intercept \( \alpha_0 \). The combination part, \( \alpha_A Y_A + \alpha_B Y_B + \ldots \), is still estimated correctly
- not well suited to high dimensional data
• combination from logistic regression is
  \[ 1.03\log(CA - 19 - 9) + 0.93\log(CA - 125) \]
Regression Trees

• finds subgroups on the basis of \{Y_A, Y_B, \ldots\} that have similar risk

• the observed proportion diseased in subgroup \( \hat{P}[D = 1|Y_A, Y_B, \ldots] \)

• each terminal node defines a subgroup by the branches that lead to it

  e.g., \((Y_A < C_A)\) and \((Y_B > C_B)\) and \((Y_C < C_C)\) \ldots

• nonparametric regression, CART

• not well suited to high dimensional data
Logic Regression

• similar but more flexible Boolean algebra

  e.g., \((Y_A < C_A)\) or \((Y_B > C_B\) and \(Y_C < C_C)\) \ldots

• can include these Boolean combination variables in a regression

  e.g., \(\log P[D = 1]/(1−P[D = 1]) = \alpha_0 + \alpha_1 B_1 + \alpha_2 B_2 + \cdots + \beta_X X\)

  where \(B_k\) is a Boolean combination and \(X\) are other predictors

• software in R

• handles high dimensional data
Other Regression Methods

• neural networks
  support vector machines
  boosting algorithms
  extreme regression

• for any technique there are data for which it is the best technique

• general recommendations hard to justify
Assessing the performance of the combination

*Principle*
separate development from performance assessment

- a training dataset to develop combination
- a test dataset to assess performance
Cross-validation

- a statistical technique to use the same data for training and testing

- leave one-out cross validation
  (i) leave out observation \((D_i, Y_i)\)
  (ii) develop model with all remaining data
    
    \[
    e.g., \beta^{(-i)}_A Y_A + \beta^{(-i)}_B Y_B + \ldots = \text{risk calculator}(-i)
    \]
  (iii) calculate \(\hat{P}^{(-i)}(D_i = 1|Y_i)\) using the risk calculator \((-i)\)
  (iv) having done (i)-(iii) for all \(n\) observations calculate ROC curve with \((D_i, \hat{P}^{(-i)}(D_i = 1|Y_i))\) \(i = 1, \ldots n\)

- 10 fold cross-validation leaves 10% of data out instead of 1 observation at a time
Cross-validation

- *all* steps in model development must be repeated for the calculator ($-i$)
- including variable selection, choosing transformations, . . . .
- having separate training and test datasets is more realistic, allows more flexibility in development
Increment in ROC performance

Figure 2. Receiver-Operating-Characteristic Curves for Death (Panel A) and Major Cardiovascular Events (Panel B) during 5-Year Follow-up.

For each end point, curves are based on models of the prediction of risk with the use of conventional risk factors with or without biomarkers (multimarker score). Biomarkers for death were B-type natriuretic peptide, C-reactive protein, the urinary albumin-to-creatinine ratio, homocysteine, and renin. Biomarkers for major cardiovascular events were B-type natriuretic peptide and the urinary albumin-to-creatinine ratio.
<table>
<thead>
<tr>
<th>Multimarker Score</th>
<th>Death</th>
<th>Major Cardiovascular Events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>adjusted hazard ratios (95% CI)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.0 (reference group)</td>
<td>1.0 (reference group)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.34 (0.83–2.18)</td>
<td>1.54 (0.98–2.40)</td>
</tr>
<tr>
<td>High</td>
<td>4.08 (2.51–6.62)</td>
<td>1.84 (1.11–3.05)</td>
</tr>
<tr>
<td>P value for trend</td>
<td>&lt;0.001</td>
<td>0.02</td>
</tr>
</tbody>
</table>

* Hazard ratios were adjusted for age; sex; body-mass index; categories of blood pressure, total cholesterol, and high-density lipoprotein cholesterol; smoking status; presence or absence of diabetes; serum creatinine level; and presence or absence of prevalent cardiovascular disease (for the model with death).

Increment in ROC performance

- often surprisingly small, because . . . .
- need large effects for substantially improved classification (Pepe et al AJE 2004)
- new markers may be correlated with existing markers
- do not use $\Delta$AUC to quantify incremental value, rather use a more meaningful measure, e.g., $\Delta$ROC($t$)
Summary

• ‘and’ ‘or’ rules for combining binary tests

• best combinations are based on calculation of risk given marker data

• logistic regression is natural for combining markers, but many other methods exist too

• to assess performance of a combination, separate training dataset from testing dataset

• when existing clinical predictors exist may assess increment in performance gained with new marker
Bibliography


