Statistical Methods for Evaluating Biomarkers

Holly Janes
Fred Hutchinson Cancer Research Center

November 10, 2008

Biomarkers for...

Diagnosis: disease versus non-disease
Screening: early diagnosis
Prognosis: predicting outcome

Examples

- clinical signs / symptoms
- laboratory tests
- gene expression technology
- proteomics
- combinations of any of the above

How to evaluate their accuracy?

Outline

1. Measures of biomarker accuracy
2. Evaluating incremental value
3. Phases of biomarker development
4. Study design issues
5. Advanced topics
6. Software

Measures of Accuracy for Binary Markers

Classification Probabilities

D = outcome (disease)
Y = binary marker

	$D = 0$	$D = 1$
$Y = 0$	True negative	False negative
$Y = 1$	False positive	True positive

false positive fraction $=$ FPF $= P[Y = 1|D = 0]$ = 1 - specificity
true positive fraction $=$ TPF $= P[Y = 1|D = 1]$ = sensitivity

Ideal test: TPF = 1 and FPF = 0

Classification Probabilities, cont'd

- condition on disease status
- describe test accuracy
- helpful to public health researchers: should the test be used?
- helpful to individual: should I have the test?

Predictive Values

positive predictive value $=$ PPV $= P[D = 1 | Y = 1]$
negative predictive value $=$ NPV $= P[D = 0 | Y = 0]$

Ideal test: PPV = 1 and NPV = 1

- condition on test result
- require cohort study to estimate
- depend on TPF, FPF, and prevalence (ρ)

  PPV = ρTPF/(ρTPF + (1-ρ)FPF)
  NPV = (1-ρ)(1-FPF)/((1-ρ)(1-FPF) + ρ(1-TPF))
- describe predictive capacity of test
- given my test result, how likely is it that I'm diseased?

Example: Diagnosis of CAD

Y : exercise stress test
D : coronary artery disease

	$D = 0$	$D = 1$	
$Y = 0$	22.3%	14.2%	36.5%
$Y = 1$	7.8%	55.6%	63.4%
	30.1%	69.8%	100%

TPF = 0.797, FPF = 0.259, ρ = 0.698
PPV = 0.877, NPV = 0.611, τ = 0.634

- CAD detects 80% of diseased subjects and incorrectly identifies 26% of non-diseased as suspicious
- 88% of test positives and 39% of test negatives have disease

From *The Statistical Evaluation of Medical Tests for Classification and Prediction* by Margaret S. Pepe, Ph.D., Oxford University Press, 2003

Inappropriate Commonly Used Measures

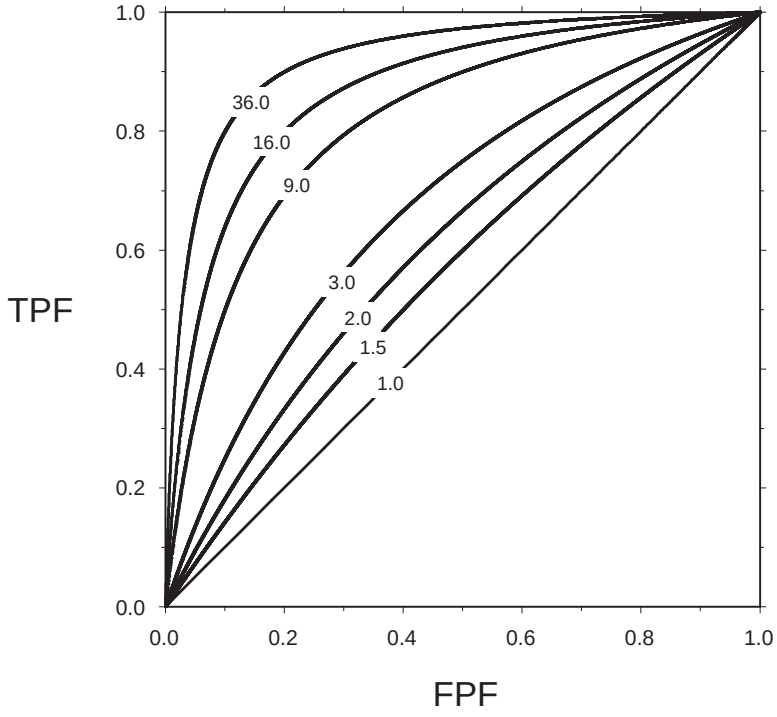
- misclassification rate (MCR)
- odds ratio

MCR

- $= P[Y \neq D]$
  $= P[Y = 1|D = 0]P[D = 0] + P[Y = 0|D = 1]P[D = 1]$
  $= \text{FPF} * (1 - \rho) + (1 - \text{TPF}) * \rho$
- ignores differential importance of false negative and false positive errors
- depends on the prevalence (ρ)
  - eg, if $P[Y = 1|D = 1] = P[Y = 1|D = 0] = 0$ with low ρ, MCR low
- used a lot in statistics, not in medical settings

Odds Ratio

- $= \frac{\text{TPF}*(1-\text{FPF})}{\text{FPF}*(1-\text{TPF})}$
- measure of association, not classification
- good classification $\Rightarrow$ huge odds ratios
- e.g., TPF $= 0.80$, FPF $= 0.10$ (a 'good' test)
  - Odds Ratio $= \frac{0.80*(1-0.10)}{0.10*(1-0.80)} = 36$



(FPF, TPF) corresponding to different odds ratios

- large odds ratio does *not* imply good classifier
- need to report FPF and TPF separately

From Pepe et al. *A*JE 2004; 159:882–90.

Measures of Accuracy for Continuous Markers

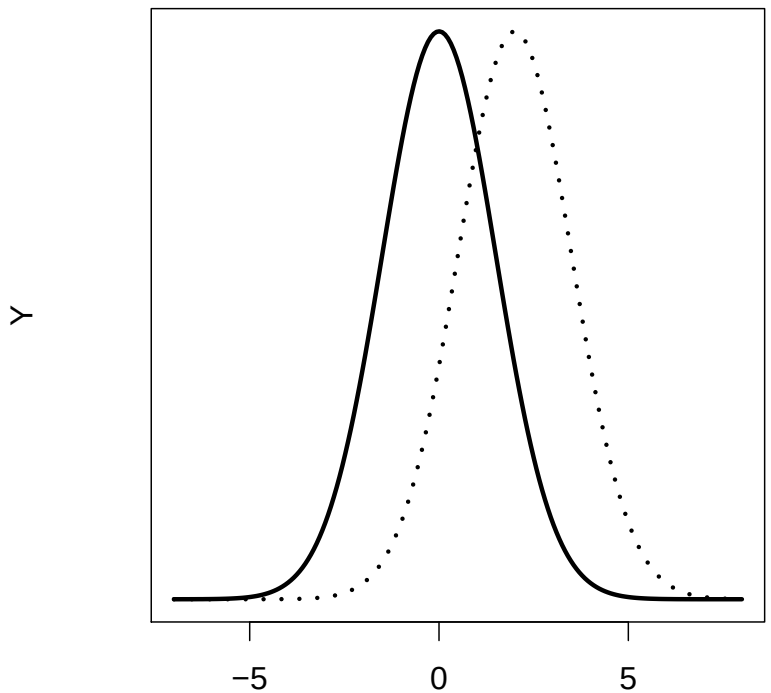
Classification Accuracy for a Continuous Test

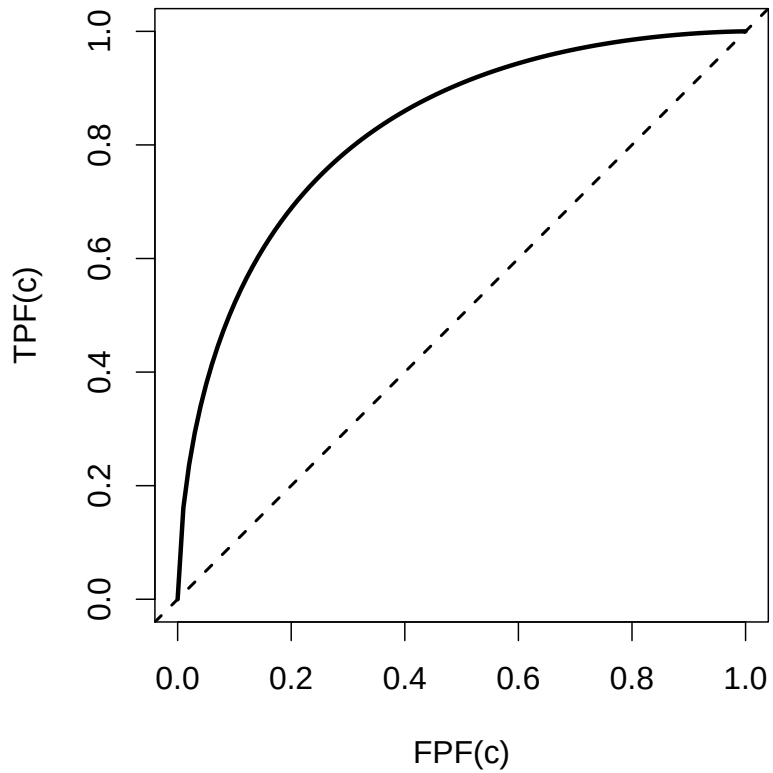
Continuous marker, Y

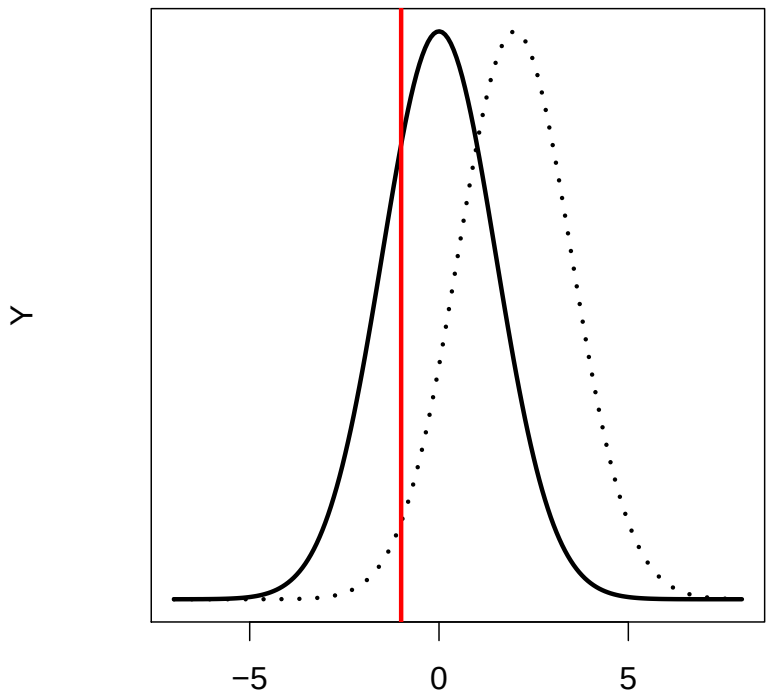
- most markers

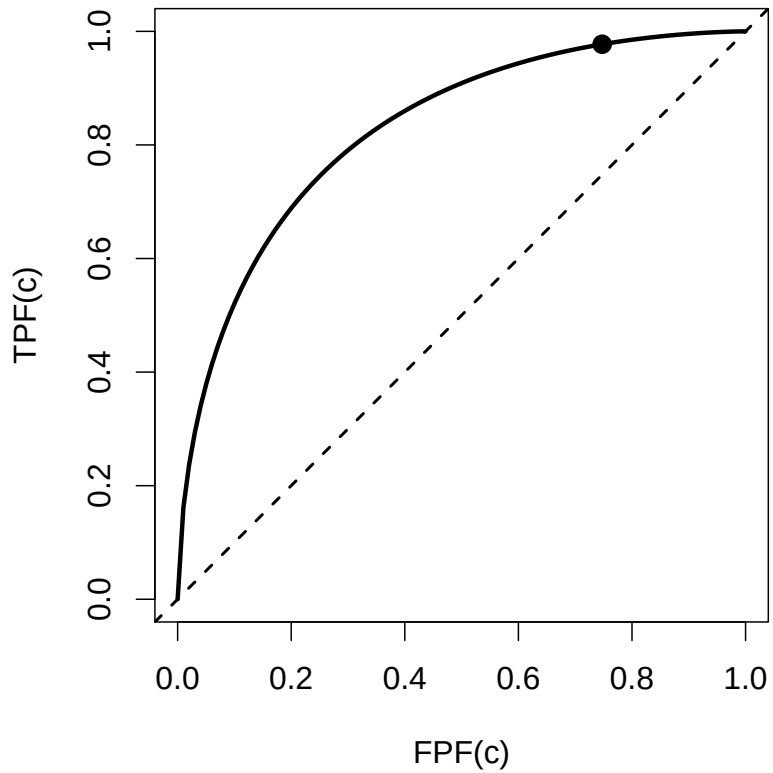
The ROC curve generalizes (FPF, TPF) to continuous markers

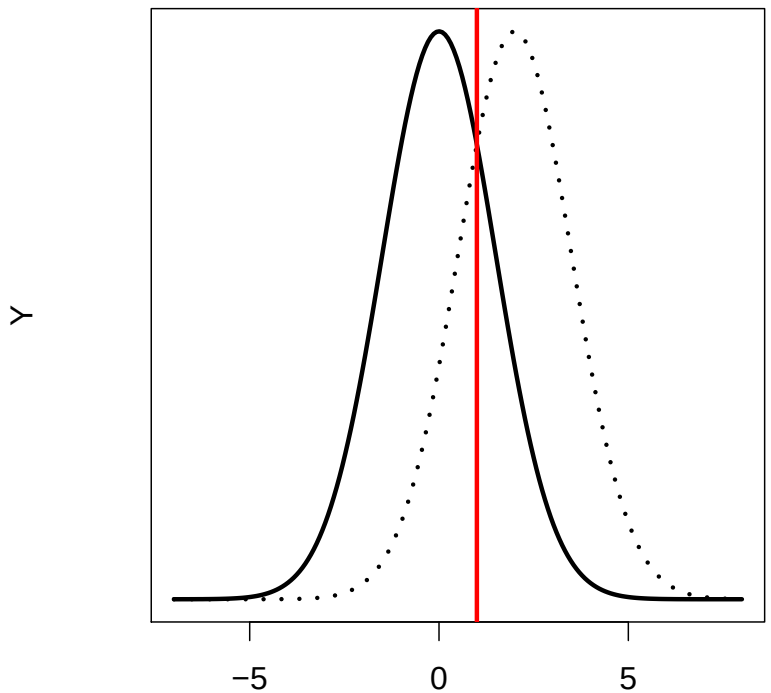
- thresholding rule: 'positive' if $Y \geq c$
- $\text{TPF}(c) = P[Y \geq c | D = 1]$
  $\text{FPF}(c) = P[Y \geq c | D = 0]$
- $\text{ROC}(\cdot) = \{(\text{FPF}(c), \text{TPF}(c)), \quad c \quad \epsilon \quad (-\infty, \infty)\}$

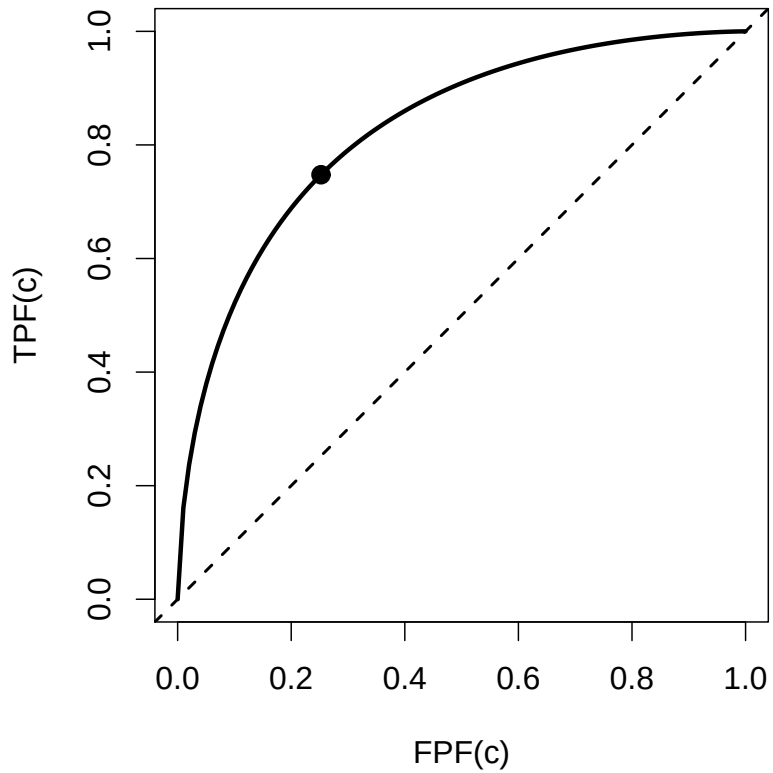

Y
−5
0
5

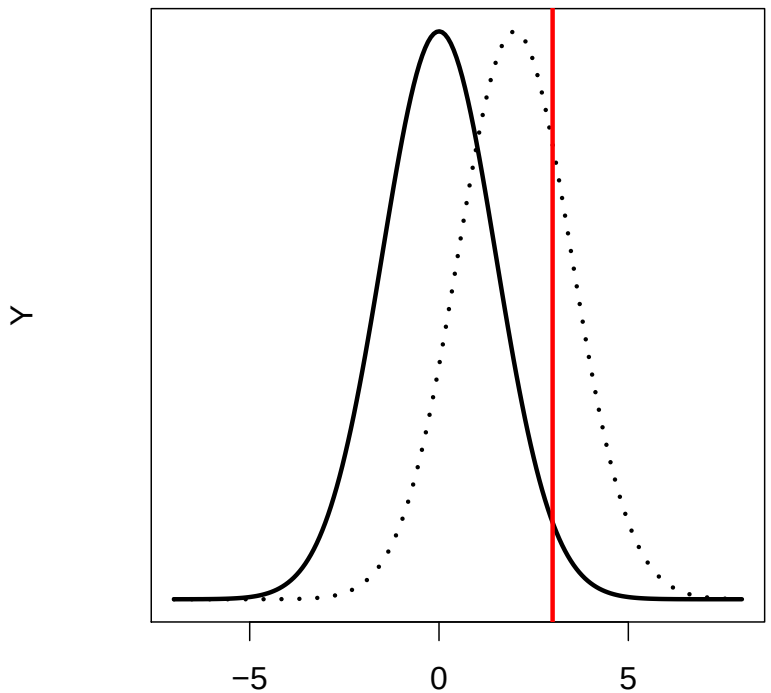

TPF(c)
0.0
0.2
0.4
0.6
0.8
1.0
FPF(c)
0.0
0.2
0.4
0.6
0.8
1.0

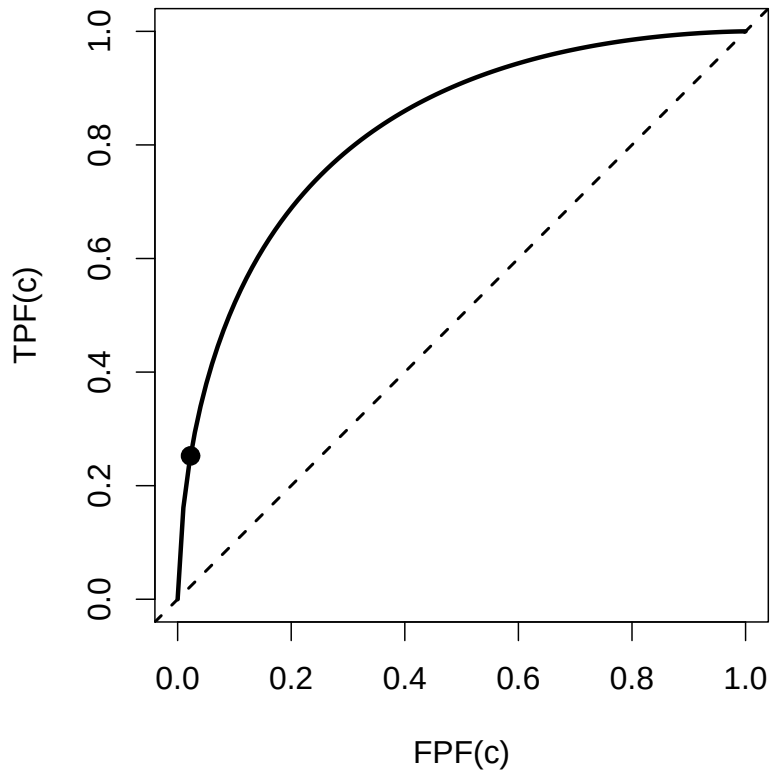

Y
−5
0
5


TPF(c)
0.0
0.2
0.4
0.6
0.8
1.0
FPF(c)
0.0
0.2
0.4
0.6
0.8
1.0


Y
−5
0
5


TPF(c)
0.0
0.2
0.4
0.6
0.8
1.0
FPF(c)
0.0
0.2
0.4
0.6
0.8
1.0


Y
−5
0
5


TPF(c)
0.0
0.2
0.4
0.6
0.8
1.0
FPF(c)
0.0
0.2
0.4
0.6
0.8
1.0

Attributes of the ROC

- shows entire range of possible performance
- puts different tests on a common relevant scale

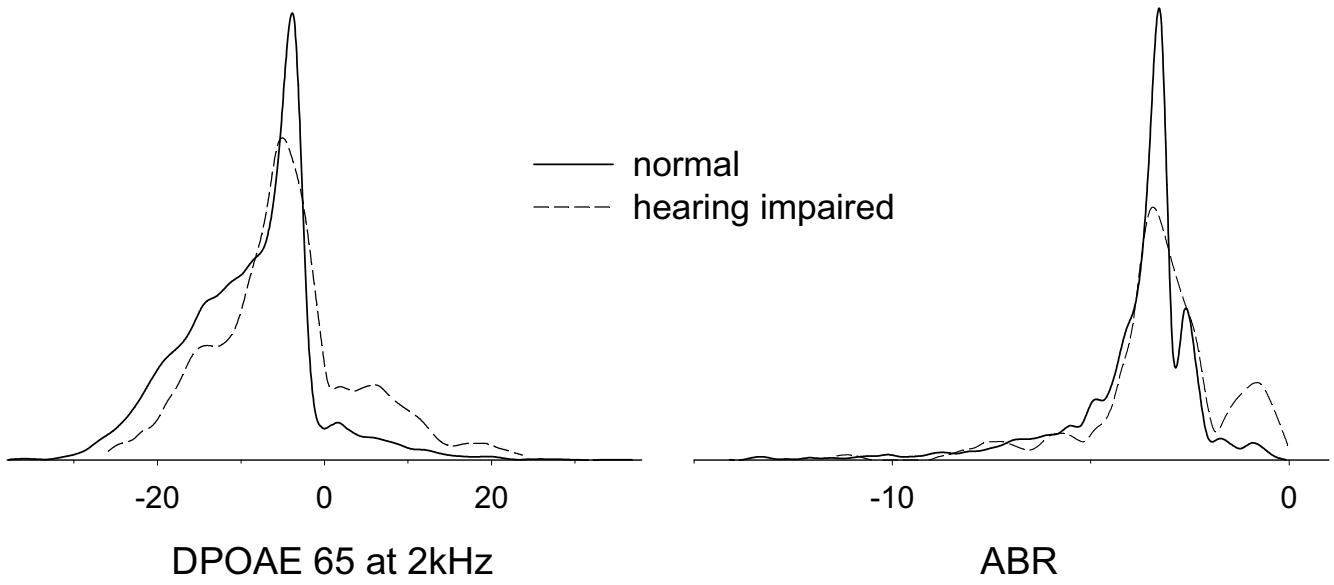


Figure 4.3 Probability distributions of test results for the DPOAE and ABR tests among hearing impaired ears and normally hearing ears.

From *The Statistical Evaluation of Medical Tests for Classification and Prediction* by Margaret S. Pepe, Ph.D., Oxford University Press, 2003

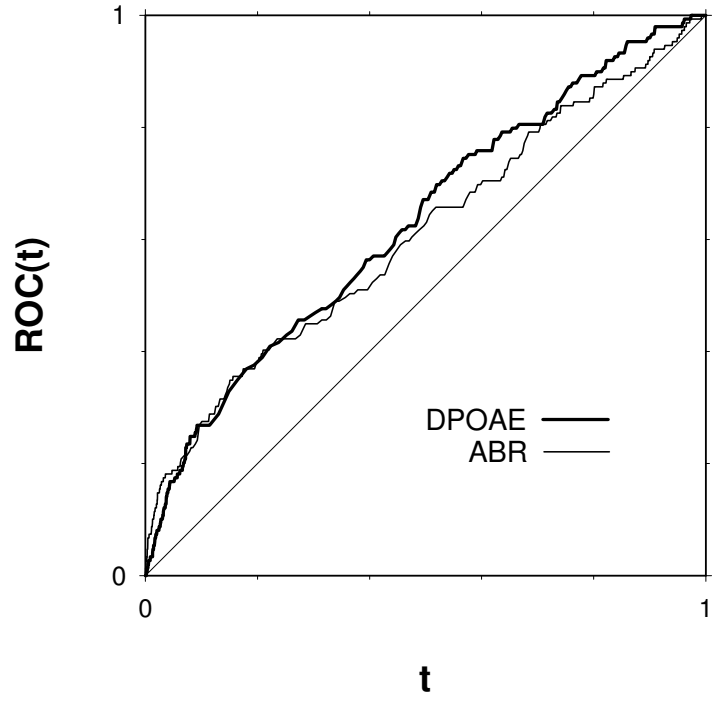


Figure 4.4 ROC curves for the DPOAE and ABR tests.

- two tests have similar ability to distinguish between hearing-impaired and normal ears

From *The Statistical Evaluation of Medical Tests for Classification and Prediction* by Margaret S. Pepe, Ph.D., Oxford University Press, 2003

Choosing a Threshold

Formal decision theory:

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

C_D is the cost of negatively classifying a diseased subject
C_N is the cost of positively classifying a non-diseased subject

$\Longrightarrow$ cost minimized at the threshold c where the slope of the ROC curve equals

$$\frac{1 - \rho}{\rho}\frac{C_N}{C_D}$$

- requires specifying costs C_D and C_N (tricky!)

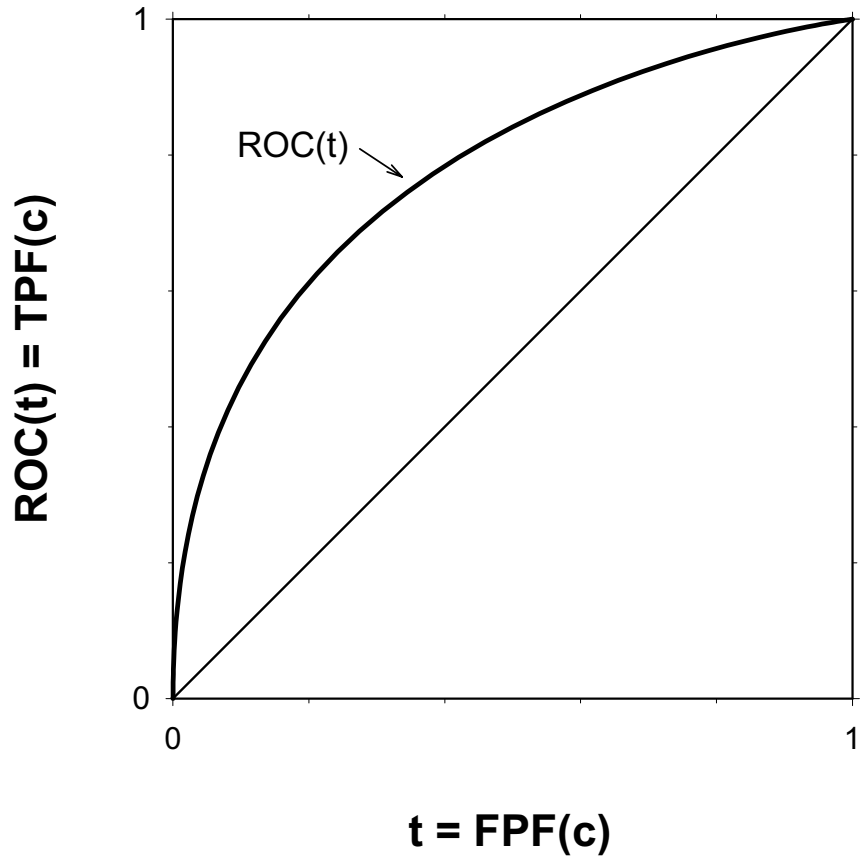
Choosing a Threshold, cont'd

Common informal practice:

- fix maximum tolerated FPF
- eg must be very low ($< 5\%$) for cancer screening test
- $f_0 = \text{FPF} \rightarrow \textit{threshold} = 1 - f_0$ quantile among controls
- or fix minimum tolerated TPF
- eg must be very high in most diagnostic settings
- $t_0 = \text{TPF} \rightarrow \textit{threshold} = 1 - t_0$ quantile among cases

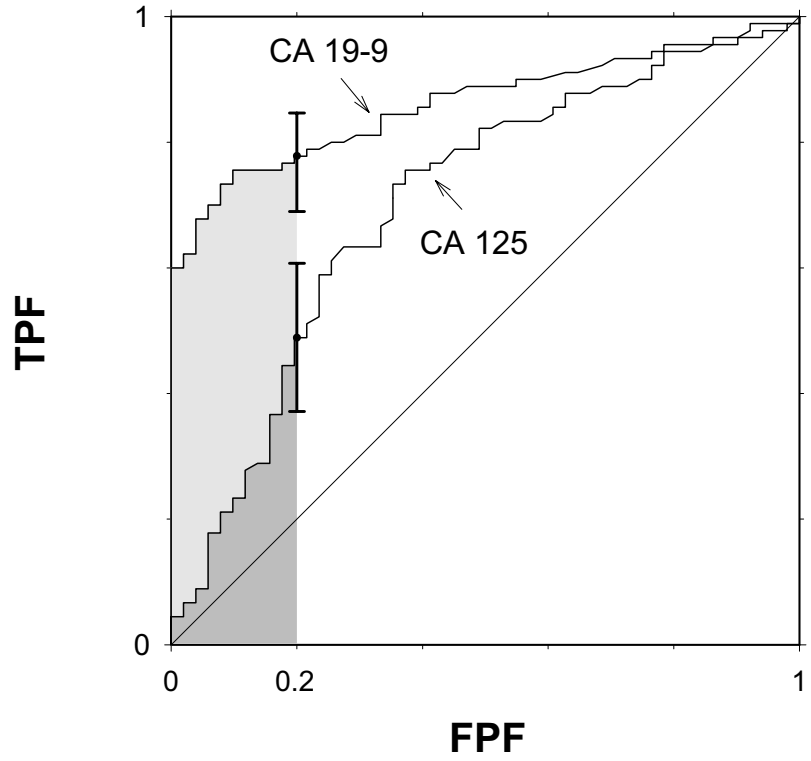
Summary Measures of Classification Accuracy

- TPF = ROC(f_0) at chosen FPF = f_0
  - percent cases detected for fixed FPF
- FPF = ROC^{-1}(t_0) at chosen TPF $= t_0$
  - FPF for fixed percent cases detected
- AUC $= \int_0^1$ ROC(f)df
  - probability of correctly ordering a randomly chosen case and control observation
  - little clinical relevance
  - summarizes TPF over entire FPF range
- partial AUC $= \int_0^{f_0}$ ROC(f_0)df
  - restricted ROC region, but little clinical relevance


1
ROC(t)
0
0
1
ROC(t) = TPF(c)
t = FPF(c)

Example: Pancreatic Cancer Data

- marker sought for screening for pancreatic cancer
- data on two markers: CA 19-9 and CA 125



From *The Statistical Evaluation of Medical Tests for Classification and Prediction*
by Margaret S. Pepe, Ph.D., Oxford University Press, 2003

AUC for CA 125 = 0.71
AUC for CA 19-9 = 0.89
p-value = 0.007
$\Longrightarrow$ the probability of correct ordering is 18% higher with CA 19-9

ROC(0.2) for CA 125 = 0.49
ROC(0.2) for CA 19-9 = 0.78
p-value = 0.04
$\Longrightarrow$ CA 19-9 detects 29% more cancers with the same FPR = 0.2

- conclusions about ROC(0.2) are more clinically important than those about AUC

Generalizing Predictive Values to Continuous Biomarkers

- a relatively new area of research; not well developed

Evaluating Incremental Value

Incremental Value

- how much classification accuracy does the new marker add to existing predictors?
- eg how much does CRP add to existing lipid measurements and risk factor information in discriminating between those who will and will not develop CVD?

How Best to Combine Markers?

- $Y = (Y_1, \ldots, Y_P)$
- the “best" combination is the risk score, $R(Y) = P(D = 1|Y_1, \ldots, Y_P)$ McIntosh and Pepe (*Biometrics*, 2000)
- “best" $\Longrightarrow$ No other combination of $(Y_1, \ldots, Y_P)$ has a (FPF, TPF) point above its ROC curve

To Combine Markers

- Estimate
$$R(Y) = P(D = 1 | Y_1, \ldots, Y_P)$$
  - using logistic regression, neural networks, classification trees, support vector machines, Bayesian modelling, ...
  - logistic regression can be used with case-control data
- Calculate the ROC curve for $R(Y)$ (it's just another marker!)
  - avoid overoptimism due to fitting and evaluating model on same data
  - split into training and validation data
  - or use cross-validation

Evaluating Incremental Value

- new marker Y^*, baseline markers $Y_1, \ldots, Y_P$
- compare the ROC curves for

$$P(D = 1 | Y_1, \ldots, Y_P)$$

and

$$P(D = 1 | Y_1, \ldots, Y_P, Y^*)$$

- NOT quantified by β^* in

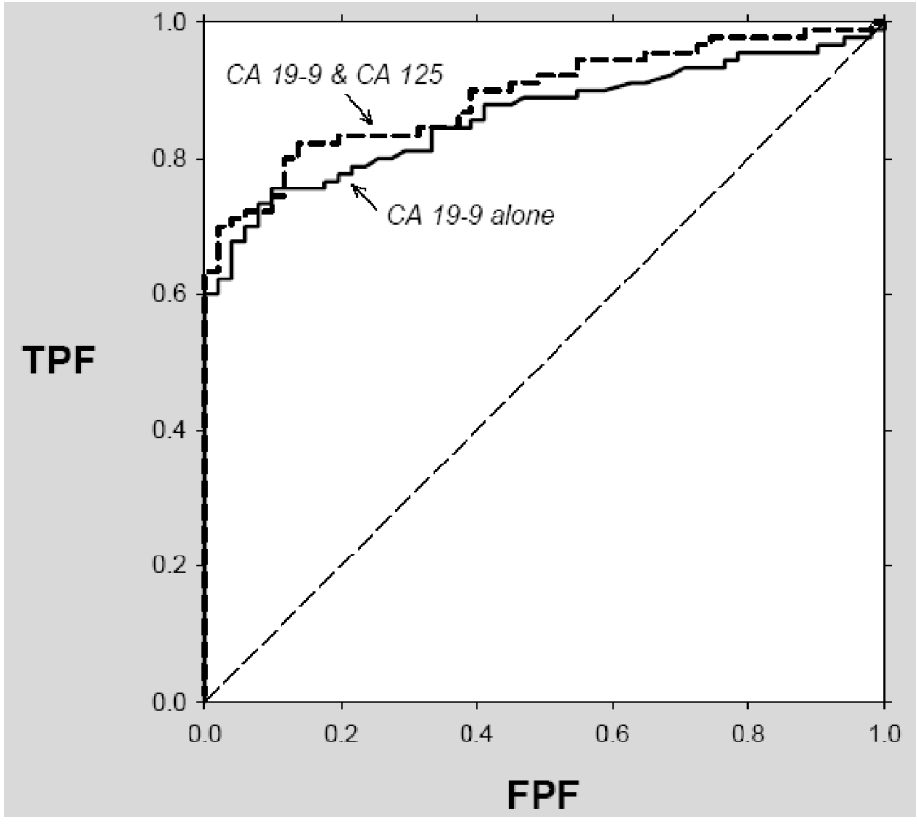
$$g(P(D = 1 | Y_1, \ldots, Y_P, Y^*)) = \beta_0 + \beta_1 Y_1 + \ldots + \beta_P Y_P + \beta^* Y^*$$

Pancreatic Cancer Example

- $Y_1 = \log$ CA-19-9 $\quad Y_2 = \log$ CA-125
- combination $\beta_1 Y_1 + \beta_2 Y_2$ from fitting

$$\text{logit} P(D = 1 | Y_1, Y_2) = \alpha + \beta_1 Y_1 + \beta_2 Y_2$$
$$\exp(\beta_2) = 2.54 \; (p = 0.002)$$

- Y_2 strongly associated with D



ROC(0.05) = 0.68 for CA 19-9
ROC(0.05) = 0.71 for combination of CA 19-9 and CA 125

- extremely common phenomenon

Phases of Biomarker Development

#	Phase	Objective	Design
1	Preclinical Exploratory	promising directions identified, assess test reproducibility	diverse and convenient cases and controls
2	Clinical Assay and Validation	clinical assay detects established disease, compare test with standard of practice, assess covariate effects	population based, cases with disease, controls without disease
3	Retrospective Longitudinal	biomarker detects disease *early* before it becomes clinical (for screening markers)	case-control study nested in a longitudinal cohort
4	Prospective Screening	extent and characteristics of disease detected by the test and the false referral rate are identified	Cross-sectional cohort of *people*
5	Disease Control	impact of screening on reducing the burden of disease on the population is quantified	randomized trial (ideally)

From: Pepe et al. Phases of biomarker development for early detection of cancer. *JNCI* **93**(14):1054–61, 2001.

Study Design Issues

Matching in Case-Control Studies

- randomly sample cases
- select controls matched to cases with respect to confounders
- attempts to eliminate confounding
- eg Physicians' Health Study
  - evaluate PSA as a screening tool for prostate cancer
  - for each case select 3 controls within 1 years of age of the case
  - cases tend to be older, older subjects tend to have higher PSA $\Longrightarrow$ age is confounder
  - matching on age attempts to correct for this

Implications of Matching

- must adjust for matching covariates in analysis
  - unadjusted analysis is biased
  - more complicated analysis
- can't assess incremental value of marker over matching covariates
- tends to increase efficiency

Selected Verification

- in prospective studies, may not be possible to obtain the outcome (disease status) for all individuals
  - too expensive (cost or resources)
  - not ethical (eg biopsy)
- often biomarker value determines whether disease status is verified
  - eg, in study of PSA and DRE for prostate cancer screening, biopsy recommended if PSA $>$ 2.5 or DRE+
- selective sampling can lead to biased estimates of accuracy – “verification bias" or “work-up bias"

Implications of Selected Verification

When comparing two binary biomarkers in paired study:

- those who test negative on both tests are not needed to estimate relative TPF, FPF

When evaluating one binary biomarker:

- naive TPF,FPF are biased
- there are methods for correcting for verification bias
- all make untestable assumptions about the verification mechanism
  - verification may depend on unmeasured factors!
- lead to decreased precision of estimated TPF
- difficult to find settings with cost savings: reduction in number verified offset by increased total sample size
- avoid selected verification whenever possible

Advanced Topics

Covariate adjustment

- adjust for covariates that impact the marker distribution in controls
- eg center effects in multicenter studies
- analogous to covariate adjustment in studies of association
- the accuracy of the marker in a population with fixed covariate value

ROC regression

- model covariate effects on biomarker accuracy
- eg disease severity
- fit regression model for ROC curve, as function of covariates

Time-dependent ROC curves

- model biomarker accuracy as a function of time between marker measurement and disease
- eg the accuracy of PSA may decline with increasing time lag between sample collection and disease
- define time-dependent versions of TPF,FPF
- model accuracy as a function of time

Imperfect reference test

- account for lack of gold standard for D
- eg questionnaire to diagnose depression
- various statistical approaches ... but is this a statistical problem?

Software

On DABS Center website: http://www.fhcrc.org/labs/pepe/dabs

- Stata packages for ROC analysis and sample size calculations by Pepe et al.
- R programs for time-dependent ROC curves by Patrick Heagerty

Websites

`http://www.fhcrc.org/labs/pepe/dabs`
DABS Center website. Contains datasets, software, references...

`http://faculty.washington.edu/~azhou/books/software.doc`
Lists some free and commercial computer programs. Also available through the Wiley website for *Statistical Methods in Diagnostic Medicine* by Zhou, Obuchowski and McClish, 2002.

`http://xray.bsd.uchicago.edu/krl/roc_soft.htm`
Charles Metz and colleagues at University of Chicago are pioneers in ROC analysis software. Developed with a focus on applications in radiologic imaging.

References

Study design

- Baker SG, Kramer BS, McIntosh M, Patterson BH, Shyr Y, Skates S. Evaluating markers for the early detection of cancer: overview of study designs and methods. *Clinical Trials* **3**:43-56, 2006.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet, HCW. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Annals of Internal Medicine* **138**:40-44, 2003.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet, HCW, Lijmer JG. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine* **138**(1):W1-12, 2003.
- Janes H, Pepe MS. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. *Biostatistics* **7**(3):456-68, 2006.

- Pepe MS, Etzioni R, Feng Z Potter JD, Thompson M, Thornquist M, Winget M and Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**(14):1054–61, 2001.
- Zhou, SH, McClish, DK, and Obuchowski, NA. Statistical Methods in Diagnostic Medicine. Wiley Press, 2002.

Combining markers

- McIntosh MS and Pepe MS. Combining several screening tests: Optimality of the risk score. *Biometrics* **58**:657–64, 2002.
- Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the ROC curve. *Biometrics* **62**:221–229, 2006.

Covariate adjustment

- Janes H, Pepe MS. Matching in studies of classification accuracy: Implications for analysis, efficiency, and assessment of incremental value. *Biometrics* 2008; 64: 1-9.

- Janes H, Pepe MS. Adjusting for Covariate Effects on Classification Accuracy Using the Covariate-Adjusted ROC Curve. *Biometrika* (in press)

ROC regression

- Alonzo TA and Pepe MS. Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**:421–32, 2002.
- Cai T and Pepe MS. Semi-parametric ROC analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* **97**: 1099–1107, 2002.
- Cai T. Semiparametric ROC regression analysis. *Biostatistics* **5**(1):45–60, 2004.
- Dodd L, Pepe MS. Partial AUC estimation and regression. *Biometrics* **59**:614–623, 2003.
- Dodd L, Pepe MS. Semi-parametric regression for the area under the Receiver Operating Characteristic Curve. *Journal of the American Statistical Association* **98**:409–417, 2003.

- Heagerty PJ and Pepe MS. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics* **48**:533–51, 1999.

Time-dependent ROCs

- Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS. The sensitivity and specificity of markers for event times *Biostatistics* **7**:182–197, 2006.
- Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* **61**:92–105, 2005.
- Zheng Y, Heagerty PJ. Semi-parametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**:615–632, 2004.

Imperfect reference test

- Albert PS, Dodd LE. A cautionary note on robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60**:427–35, 2004.

- Albert PS, McShane LM, Shih JH, et al. Latent class modeling approaches for assessing diagnostic error without a gold standard. *Biometrics* **57**:610–19, 2001.
- Alonzo TA and Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* **18**:2897-3003, 1999.
- Pepe MS, Janes H. Insights into latent class analysis. *Biostatistics* **8**:474-84, 2007.
- Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41**:959-68, 1985.

Verification bias

- Alonzo TA. Verification bias-corrected estimators of the relative true and false positive rates of two binary screening tests. *Statistics in Medicine* **24**:403–417, 2005.
- Alonzo TA and Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Applied Statistics* **54**:173–190, 2005.

- Alonzo TA, Braun TB, Moskowitz CS. Small sample estimation of relative accuracy for binary screening tests. *Statistics in Medicine* **23**:21–34, 2004.
- Alonzo TA, Kittelson JC. A novel design for estimating relative accuracy of screening tests when complete disease verification is not feasible. *Biometrics* DOI: 10.1111/j.1541-0420.2005.00445.x (early online access).
- Begg CB and Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **39**:207-15, 1983.
- Kosinski AS and Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* **59**:163-71, 2003.
- Obuchowski NA, Zhou X. Prospective studies of diagnostic test accuracy when disease prevalence is low. *Biostatistics* **3**:477-92, 2002.
- Pepe MS and Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics* **2**:1–12, 2001.

- Pepe MS and Alonzo TA. Reply to Letter to Editor regarding Alonzo TA and Pepe MS, Assessing the accuracy of a new diagnostic test when a gold standard does not exist. *Statistics in Medicine* **20**:656–660, 2001.
- Punglia et al. Effect of verification bias on screening for prostate cancer by measurement of PSA. *NEJM* **349**:335-42, 2003.