Receiver Operating Characteristics (ROC) Analysis and Sample Size Calculation

Kai Ding, PhD
Associate Professor, Biostatistics
President's Associates Presidential Professor
Department of Biostatistics and Epidemiology
Hudson College of Public Health
kai-ding@ouhsc.edu

Invited COBRE Lecture
December 2, 2022

Outline

- Background and Motivation
- ROC Analysis
- Sample Size Calculation

Background and Motivation

Study Design

- **Diagnostic Studies**: Cross-sectional, patients suspected of having a particular disease

- **Prognostic studies:** Cohort (prospective preferred), patients at risk of the outcome

Diagnostic Studies

- Diagnostic tests – goal is to distinguish between those with target disease and those without in patients suspected of having a particular disease

- ***Multivariable nature*** of the diagnostic process
  - Diagnostic determinants – findings from history, physical exam, dx test results
  - Objectives
    - Evaluate individual test accuracy
    - ID combination(s) of tests that have the largest diagnostic yield
    - Does new test provide additional diagnostic value in clinical practice?
    - Is a less burdensome or inexpensive test an alternative?

Prognostic Studies

- **Goal: individual risk prediction**
  - Gain knowledge about the occurrence of future outcomes given ***combinations*** of prognostic predictors.
- multivariable approach in design and analysis
- End product: outcome probabilities and predictive tools

- Objectives of prognostic research
  - **Which combination** of determinants **best predicts** the future outcome?
  - What is the **additional predictive value beyond** other available predictors?
  - may include **comparison of the predictive accuracy** of two (new) markers.

Diagnostic/Prognostic Test Accuracy

- Diagnostic research outcomes typically dichotomous
- Prognostic research outcomes also typically dichotomous but may comprise continuous variables such as tumor growth, pain scale, etc.
- Quantifying Test Accuracy
  - Diagnostic
    - Sensitivity and specificity
    - Predictive values
    - Likelihood ratios
    - Diagnostic Odds Ratio
    - Area under ROC curve (AUC) analysis
  - Prognostic
    - AUC analysis

Sensitivity, Specificity and Predictive Values

		Disease status		
		Has disease	No disease	Total
Test Result	Positive	A	B	A + B
	Negative	C	D	C + D
	Total	A + C	B + D	A + B + C + D

PPV=A/(A+B)
NPV=D/(C+D)

Sensitivity = A/ (A+C)

Specificity = D/ (B+D)

"Case-control" Sampling

		Disease status		
		Has disease	No disease	Total
Test Result	Positive	A	B	A + B
	Negative	C	D	C + D
	Total	A + C	B + D	A + B + C + D

~~PPV=A/(A+B)~~

~~NPV=D/(C+D)~~

Likelihood Ratios (aka as ‘Bayes Factor’)

- Likelihood ratio of a positive test: is the test more likely to be positive in diseased than non-diseased persons?

- LR+ = p(T+|D+) / p(T+|D-) = Sn/(1-Sp) = TPR / FPR
  - High LR+ values help in RULING IN the disease
  - E.g. LR+ of 10 means a diseased person is 10 times more likely to have a positive test than a non-diseased person
  - Values close to 1 indicate poor accuracy

Likelihood Ratios (aka as 'Bayes Factor')

- Likelihood ratio of a negative test: is the test less likely to be negative in the diseased than non-diseased persons?

- LR- = p(T-|D+) / p(T-|D-) = (1-Sn)/Sp = FNR / TNR
  - Low LR- values help in RULING OUT the disease
  - E.g. LR- of 0.5 means a diseased person is half as likely to have a negative test than a non-diseased person
  - Values close to 1 indicate poor accuracy

Clinical Scenario: Does This Adult Patient Have Septic Arthritis?*

A 48-year-old woman with a history of rheumatoid arthritis who has been treated with long-term, low-dose steroids presents to the emergency department with a 2-day history of a red, swollen, tender right knee.

The authors estimated the <u>pre-test</u> probability of septic arthritis is 0.38.

On examination, she is afebrile and has a fluid in her right knee joint. An arthrocentesis (needle in the joint) is done to obtain some joint fluid for analysis.

Her synovial fluid white blood cell (WBC) count is 48,000/uL.

How do you use the synovial WBC result to revise the probability of septic arthritis?

*Margaretten, M. E., J. Kohlwes, et al. (2007). <u>JAMA</u> **297**(13): 1478-88.

Make It a Dichotomous Test

Synovial WBC Count	Septic Arthritis Yes	No
>25,000	77%	27%
≤ 25,000	23%	73%
TOTAL*	100%	100%

*Note that these could have come from a study where the patients with septic arthritis (D+ patients) were sampled separately from those without (D- patients).

Make It a Dichotomous Test

Sensitivity = 77%

Specificity = 73%

LR(+) = 0.77/(1 - 0.73) = 2.9

LR(-) = (1 - 0.77)/0.73 = 0.32

“+” = > 25,000/uL

“-” = ≤ 25,000/uL

Clinical Scenario
Synovial WBC = 48,000/uL

- Pre-test probability of disease: 0.38
- Pre-test odds of disease: 0.38/0.62 = 0.61
- LR(+) = 2.9 (where > 25,000/uL =“+”)
- By a formula: Post-Test Odds (given the “+” test) = Pre-Test Odds × LR(+) = 0.61 × 2.9 = 1.75
- **Post-Test probability of disease = 1.75/(1.75+1) = 0.64**

Clinical Scenario
Synovial WBC = 128,000/uL

- Pre-test probability of disease : 0.38
- Pre-test odds of disease: 0.38/0.62 = 0.61
- LR(+) = 2.9 (where > 25,000/uL ="+")
  - same as for WBC=48,000!
- By a formula: Post-Test Odds (given the "+" test) = Pre-Test Odds × LR(+) = 0.61 × 2.9 = 1.75
- **Post-Test probability of disease = 1.75/(1.75+1) = 0.64**

Same post-test probability although test result are more positive!!

Criterion of Test Positivity: Impact on Sn and Sp

- Sensitivity and specificity depend on the cut-point chosen to separate test “positives” from test “negatives”.
- High threshold → few false positives (higher specificity) but many false negatives (lower sensitivity)
- Low threshold → more false positives (lower specificity) but fewer false negatives (higher sensitivity)

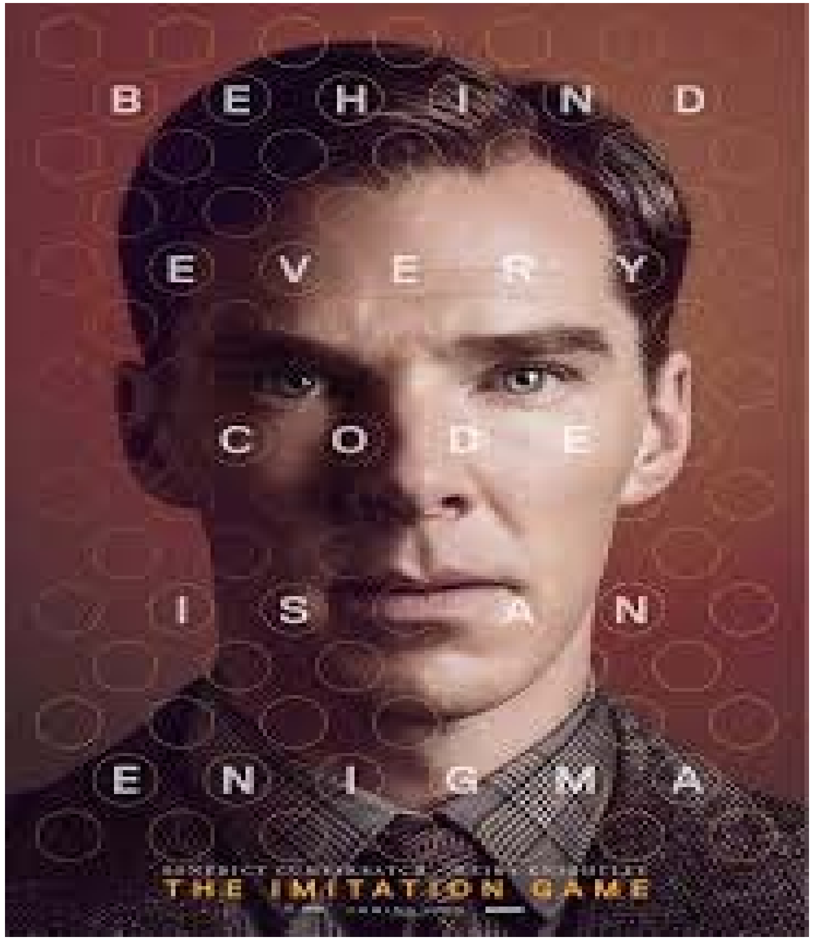
Implications of Choice of Cut-point

- For **tests measured on ordinal or continuous scales**, a single cut-off value does not fully characterize test performance
  - In this example, we regard probability of joint infection as equal whether the synovial WBC count is 48,000/uL or 128,000/uL.

- We need a flexible way to understand the performance of a continuous/ordinal test that permits use of multiple cut-points: **the receiver operating characteristic (ROC) curve**

ROC Analysis

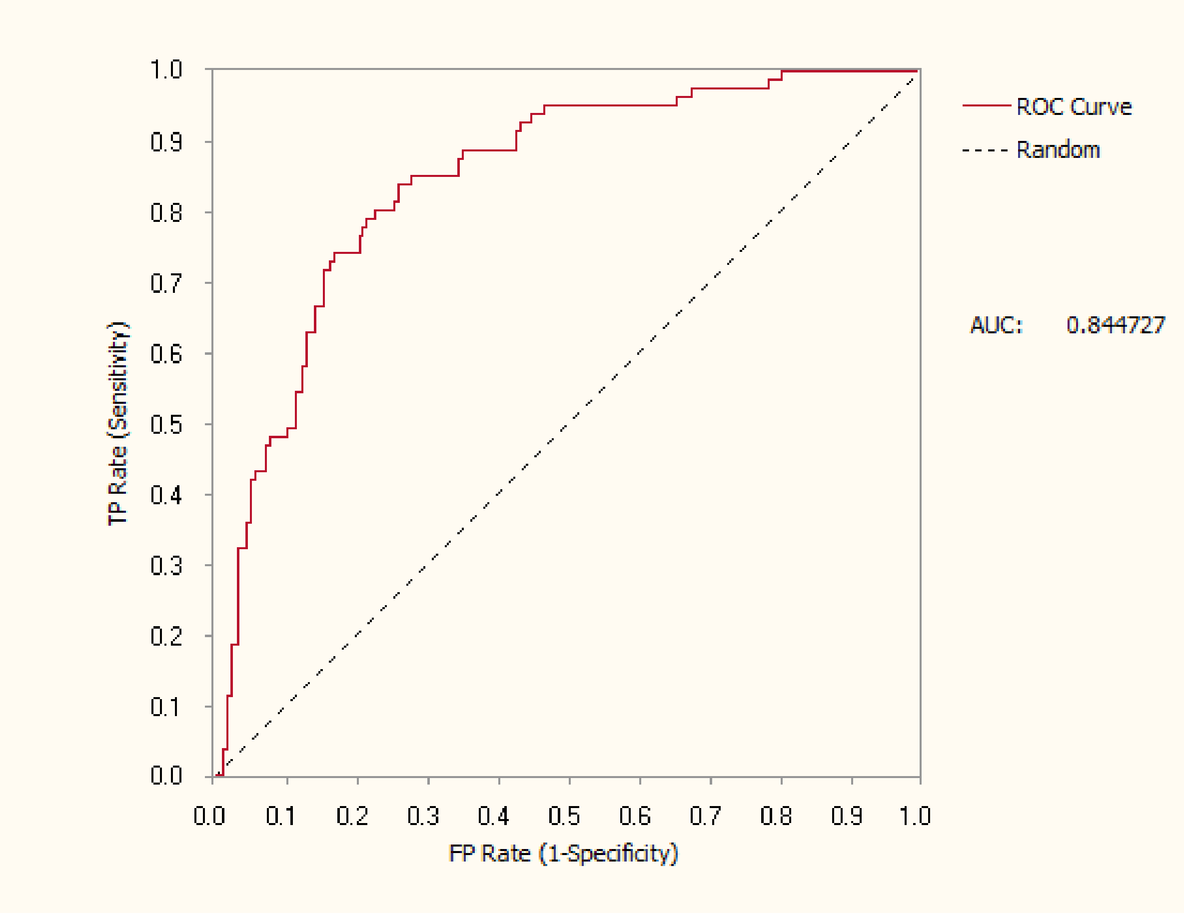
ROC Curve: A Brief History



- Part of a field called "*Signal Detection Theory*" developed during World War II for the analysis of radar images.
- Blip on the screen - an enemy target, a friendly ship, or just noise.
- Radar receivers' ability to make these important distinctions was called the ***Receiver Operating Characteristics (ROC).***
- Used in medicine, radiology, biometrics, forecasting of natural hazards, meteorology, model performance assessment, and increasingly in machine learning and data mining research.

ROC Curve

- Illustrates sensitivity and specificity tradeoffs as we vary the cutoff point
- A plot of FP probability on the x-axis and TP probability on the y-axis across several thresholds of a continuous value
- Each point on the curve represents a Se/Sp pair corresponding to a particular cutoff (decision threshold or criterion value)
- AUC is the area between the curve and the X-axis



AUC Estimation

- **Parametric AUC (Fitted or Smooth ROC curve)** distributional assumptions
  - Test results (or some unknown monotonic transformation of them) follow a binormal distribution
  - Maximum likelihood estimation
  - Preferred method for **discrete rating data** e.g. a 5-point scale

- **Non-parametric AUC (Empirical ROC curve)**
  - Most commonly used in clinical research
  - Connect all the points obtained at all the possible cutoff levels
  - ***Summation of the areas of the trapezoids*** formed by connecting the points on the ROC curve

- For continuous or quasi-continuous data the parametric and nonparametric estimates of AUC will have very similar values

ROC Curve

- Drawing the ROC curve requires **varying** the cut-point, **not choosing** a fixed cut-point.

- The ROC curve is drawn by **serially lowering** the cut-point **from highest** (most abnormal) **to lowest** (least abnormal).

- ROC curve is for evaluating the test, not the patient
  - Not particularly useful in interpreting a test result for a given patient

Septic Arthritis Example

WBC (/uL) interval	**% of septic arthritis**	**% of no septic arthritis**
>100,000	29%	1%
50,001 – 100,000	33%	7%
25,001 – 50,000	15%	19%
0 – 25,000	23%	73%
TOTAL	**100%**	**100%**

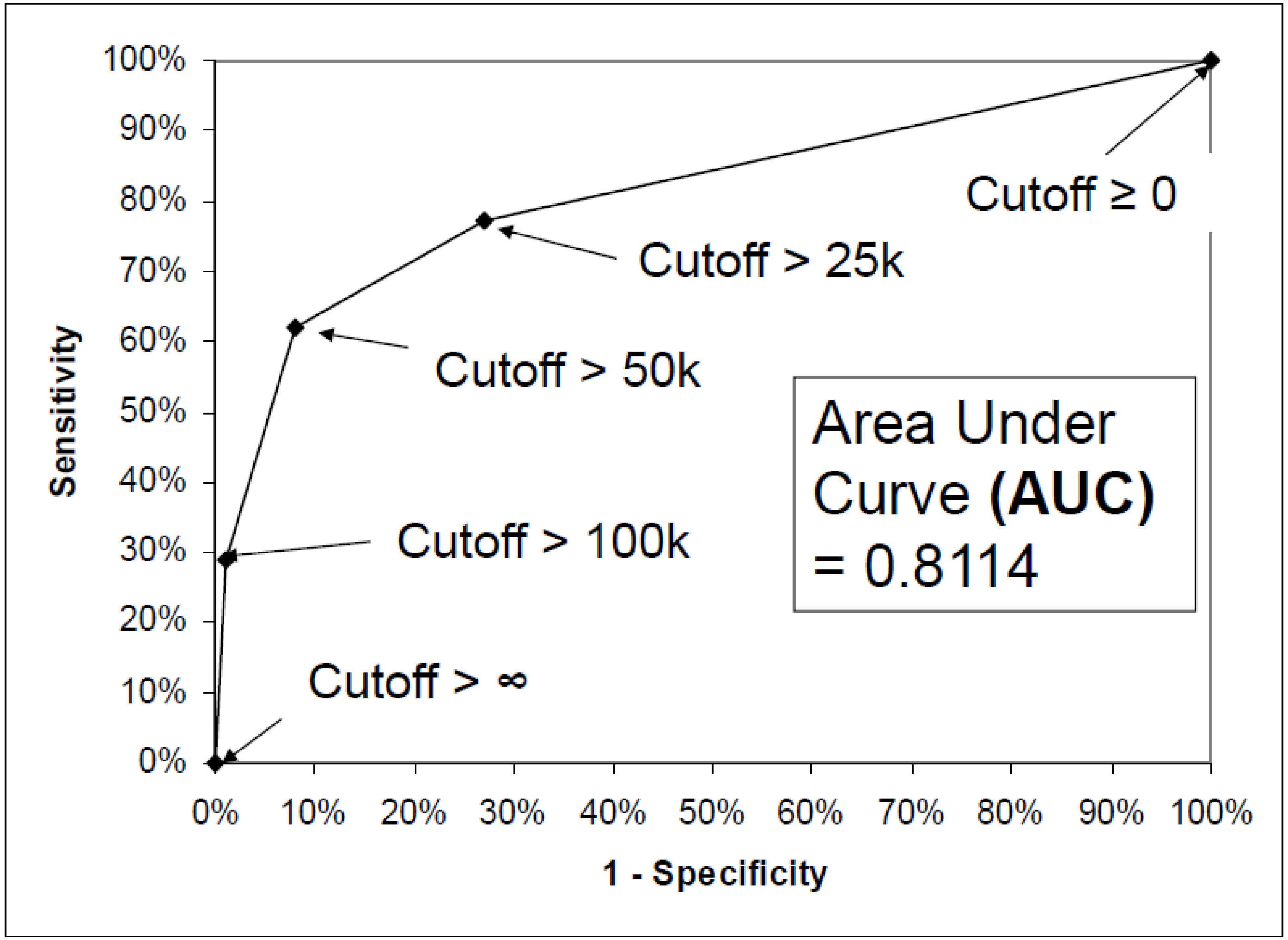
Margaretten, M. E., J. Kohlwes, et al. (2007). Jama **297**(13): 1478-88.

Convert to ROC Table

WBC Count (x1000/uL)	Sensitivity	1 - Specificity
> highest	0%	0%
> 100	29%	1%
> 50	62%	8%
> 25	77%	27%
≥ 0	100%	100%

Margaretten, M. E., J. Kohlwes, et al. (2007). Jama **297**(13): 1478-88.

ROC Curve



100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Sensitivity
Cutoff ≥ 0
Cutoff > 25k
Cutoff > 50k
Cutoff > 100k
Cutoff > ∞
Area Under Curve **(AUC)** = 0.8114
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
1 - Specificity

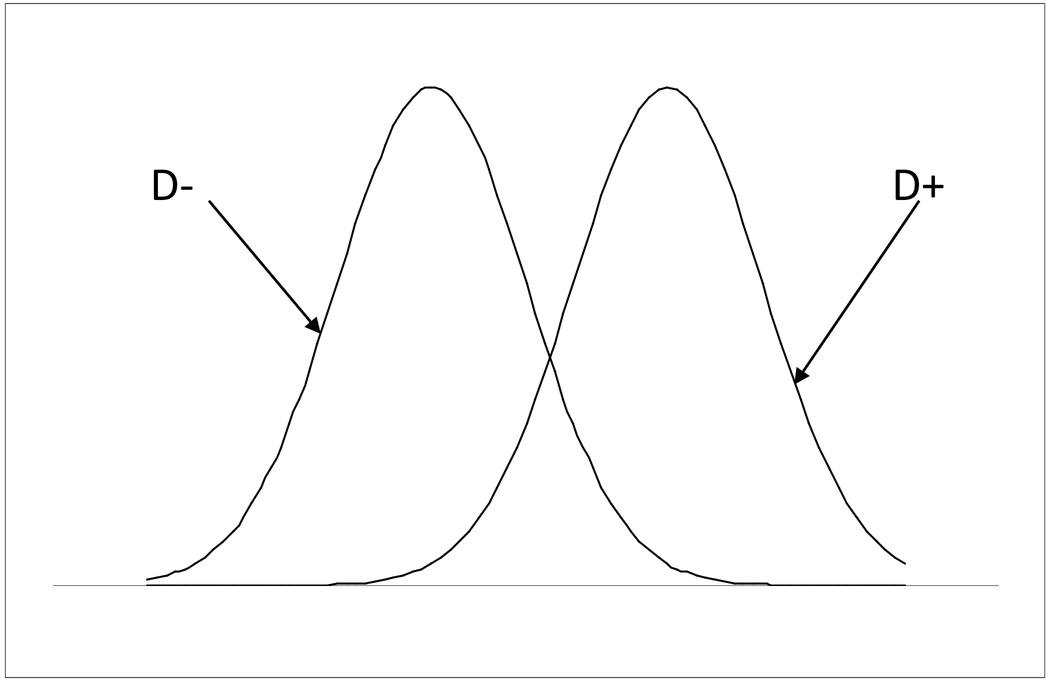
AUC Interpretation

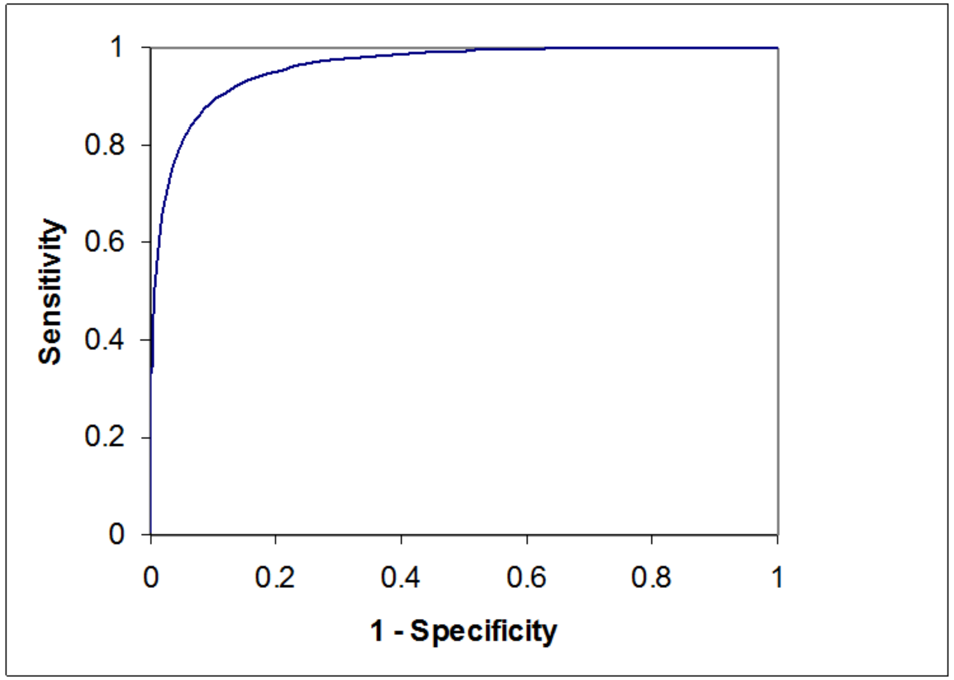
- Quantifies the discrimination of the test/predictor variable(s)
- Equals the probability that, given a pair of randomly chosen patients, one of whom truly has the outcome of interest and the other truly does not, the test will accurately identify which of the pair has the outcome.
- Equivalent to c-statistic generated by logistic regression

Accuracy	AUC
Non-informative	AUC = 0.5
Less accurate	0.5 < AUC < 0.7
Moderately accurate	0.7 < AUC < 0.9
Highly accurate	0.9 < AUC < 1
Perfect test	AUC = 1
Results for PT IgG	
Area under the ROC curve (AUC)	0.798
Standard error[a]	0.0177
95 % confidence interval[b]	0.763–0.832
Z statistic	16.836
Significance level P (area = 0.5)	<0.0001

[a]Hanley and McNeil (1982)
[b]AUC ± 1.96 SE

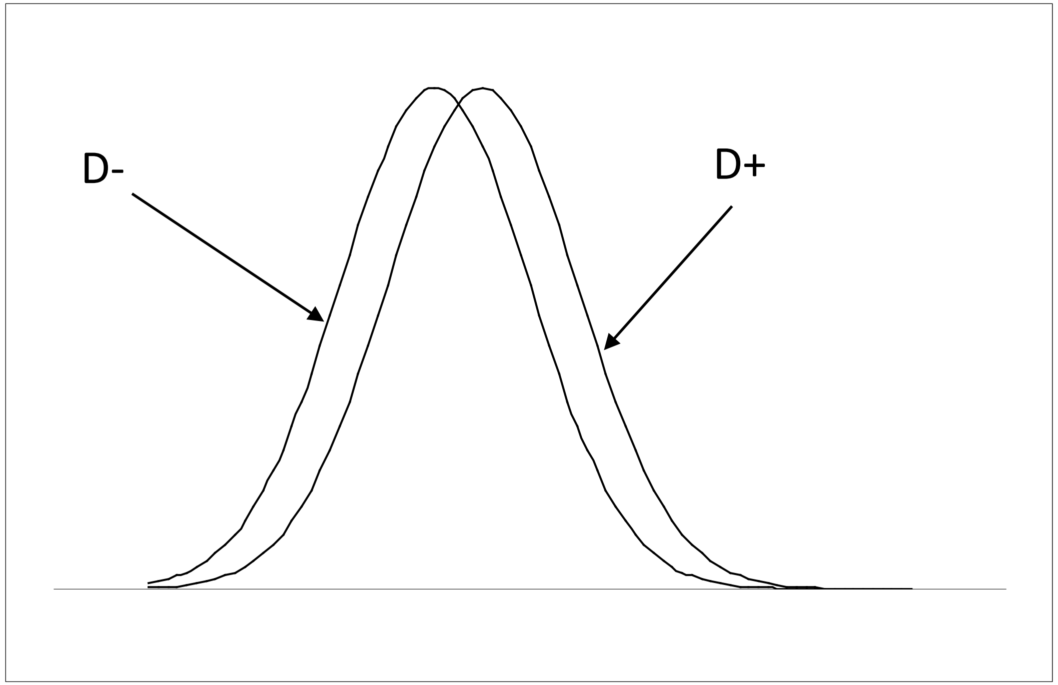
Test Discriminates Fairly Well Between D+ and D-



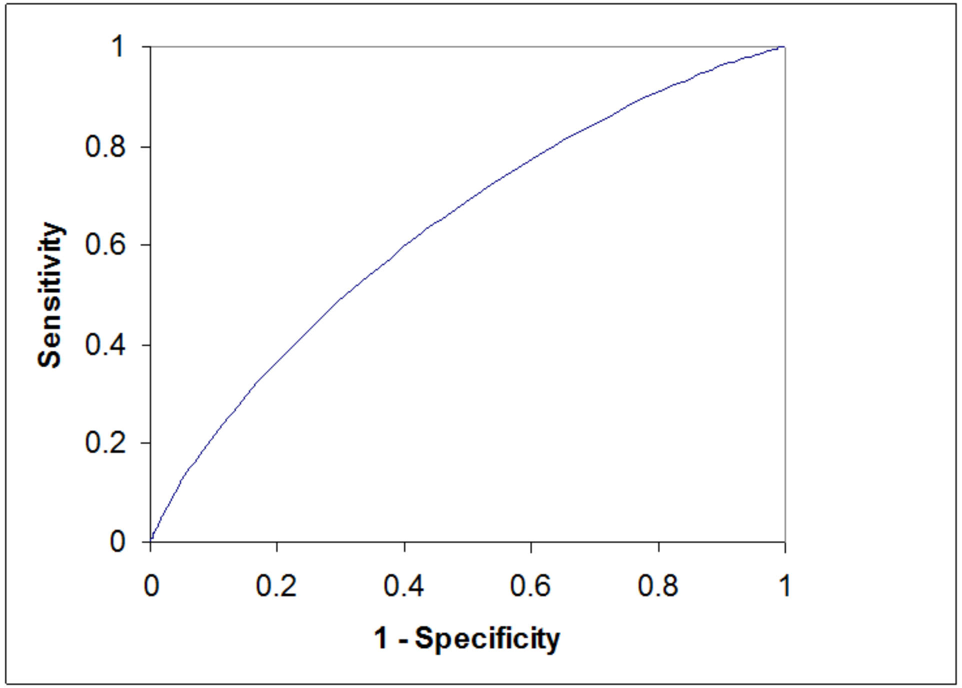


Test Result

Test Discriminates Poorly Between D+ and D-



Test Result



Summary of ROC Uses in Clinical Research

- In clinical practice, an adequate diagnosis, prediction of the course of an illness are major daily concerns.

- ROC can be used to
  - Evaluate test performance (predictive accuracy of test/prognostic factor)
    - External validation of diagnostic and prognostic models
  - Compare the predictive performance of two or more tests/factors
    - Added diagnostic/prognostic value
  - Select threshold/cut-point

Test Performance and Comparison of Two or More Tests

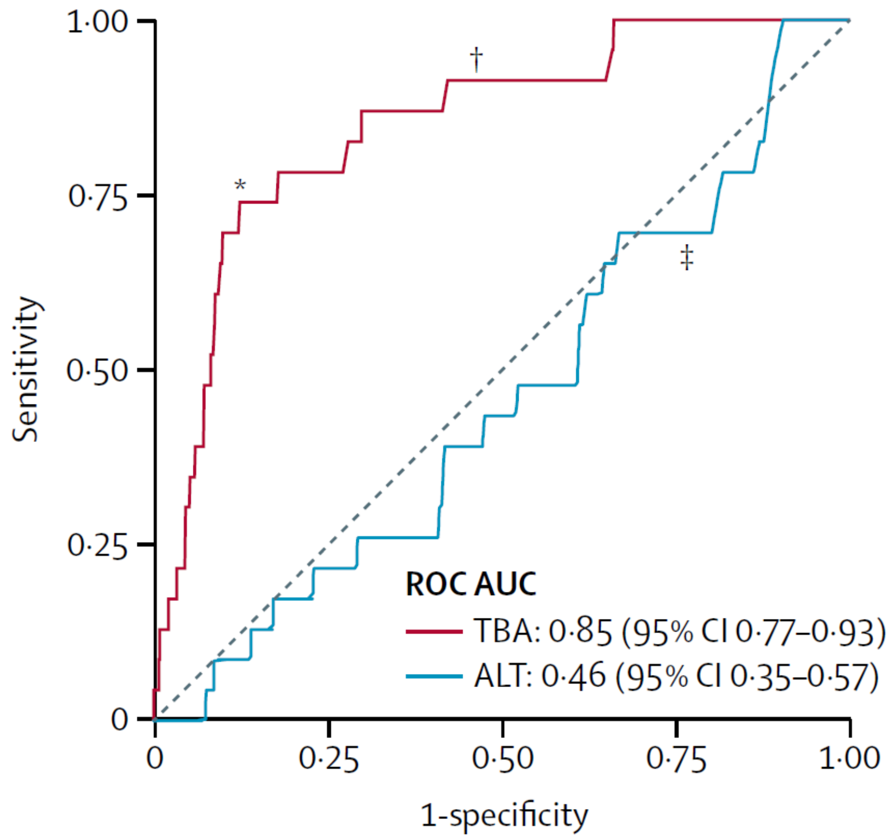


Figure 3: ROC curves for the association between stillbirth and serum biochemical markers for singleton Pregnancies

Ovadia et al. (2019) The Lancet, 393(10174), 899-909

Comparing Two or More Tests

- In some cases, AUC values can be equal, which means that the two tests yield the same overall diagnostic performance.
  - Shape of the ROC curves with equal AUC may not be identical.

- In some instances, only a small portion of the ROC curve may be of interest when comparing 2 diagnostic tests.
  - Comparing the AUCs and the overall diagnostic performance may be misleading.

Comparing Two or More Tests

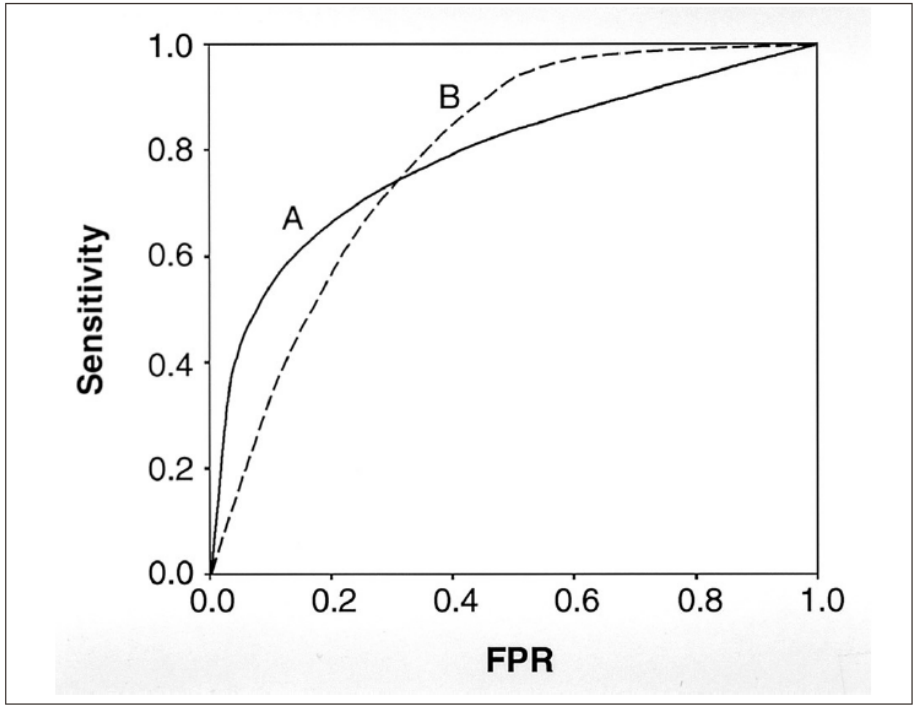


Fig. 3. Two ROC curves (A and B) with equal area under the ROC curve. However, these two ROC curves are not identical. In the high false positive rate range (or high sensitivity range) test B is better than test A, whereas in the low false positive rate range (or low sensitivity range) test A is better than test B.

Park et al, Korean J Radiol, 2004

Sensitivity at a Particular FPR and Partial Area Under the ROC Curve (pAUC)

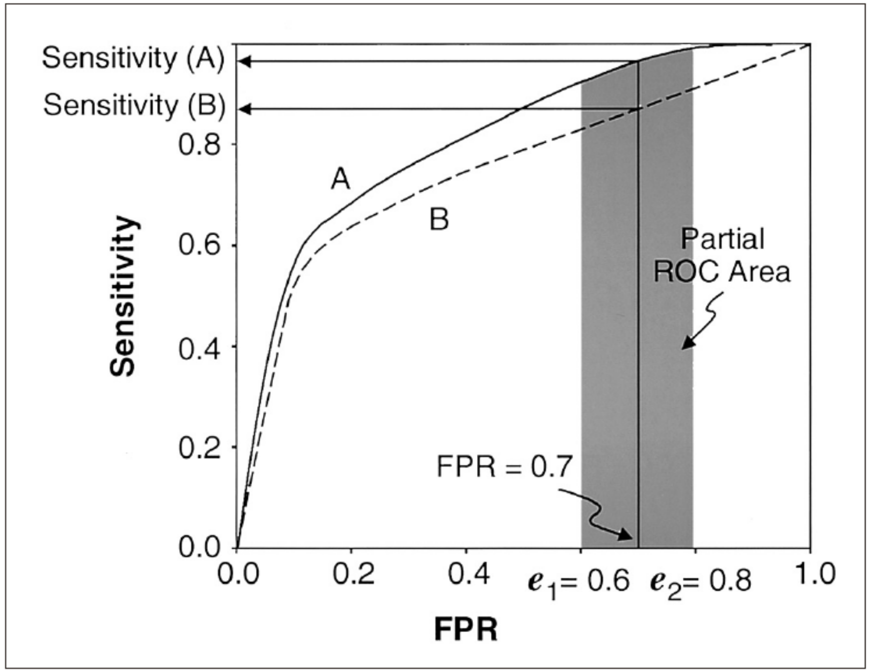


Fig. 4. Schematic illustration of a comparison between the sensitivities of two ROC curves (A and B) at a particular false positive rate and comparison between two partial ROC areas. For this example, the false positive rate and partial range of false positive rate ($e_1 - e_2$) are arbitrarily chosen as 0.7 and 0.6 ~ 0.8, respectively.

Threshold Selection

- Mathematical criteria
  - Maximum absolute sum of Sn and Sp
  - Youden Index (J): Maximum (Sn + Sp – 1)
- Clinical criteria
  - Variable depending on condition under study
  - May favor sensitivity over specificity or the other way around
- Cost Minimization/Decision-Making criteria
  - Considers the financial cost, health impact, discomfort to patient and further investigative cost (downstream cost) for correct and false diagnosis. Also <u>factors in prior probability of disease</u>
  - Sn and Sp
  - Likelihood ratio

Cost-Minimization Criterion

- The optimal cut-point, from the decision-making criterion, depends on
  - The pre-test probability of disease
  - The relative cost of failing to treat (B) vs. the cost of treating unnecessarily (C)
    - B=False negative cost; C=False positive cost
    - Misclassification cost ratio (MCR) = C/B, also called threshold odds
    - Expected MCR=(C/B)*(1-P)/P, where P=prior probability of disease

Decision Making/Cost Minimization Criterion: Using Sn and Sp

- Maximize the function: ***Sensitivity –m(1-Specificity),*** where

$$m = \left(\frac{false - positive\ cost}{false - negative\ cost} \right) \left(\frac{1 - P}{P} \right)$$

Zweig, MH, Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 1993;39/4, 56-577.

Decision Making/Cost Minimization Criterion: Using Likelihood Ratio

- The optimal cut-point, from the decision-making criterion, depends on
  - Slope of the ROC curve (i.e., likelihood ratio of certain type)
  - Relative cost of failing to treat (B) vs. cost of treating unnecessarily (C)
  - Pre-test probability of disease

Treatment Threshold Probability (PTT)

- First introduced by Pauker and Kassirer in 1975
- It is the probability of disease at which the **expected costs** of the two types of mistakes we can make (treating people without the disease **(C)** and not treating people with the disease **(B)**) **are balanced**.
- Expected cost = multiply the cost of these mistakes (C and B) by their probability of occurring.
  - The expected cost **of not treating** is P (the probability of disease) x B = **PB**
  - The expected cost **of treating** is (1 – P) (i.e., the probability of NO disease) x C= **(1 - P) x C = (C- C x P)**

X-Graph

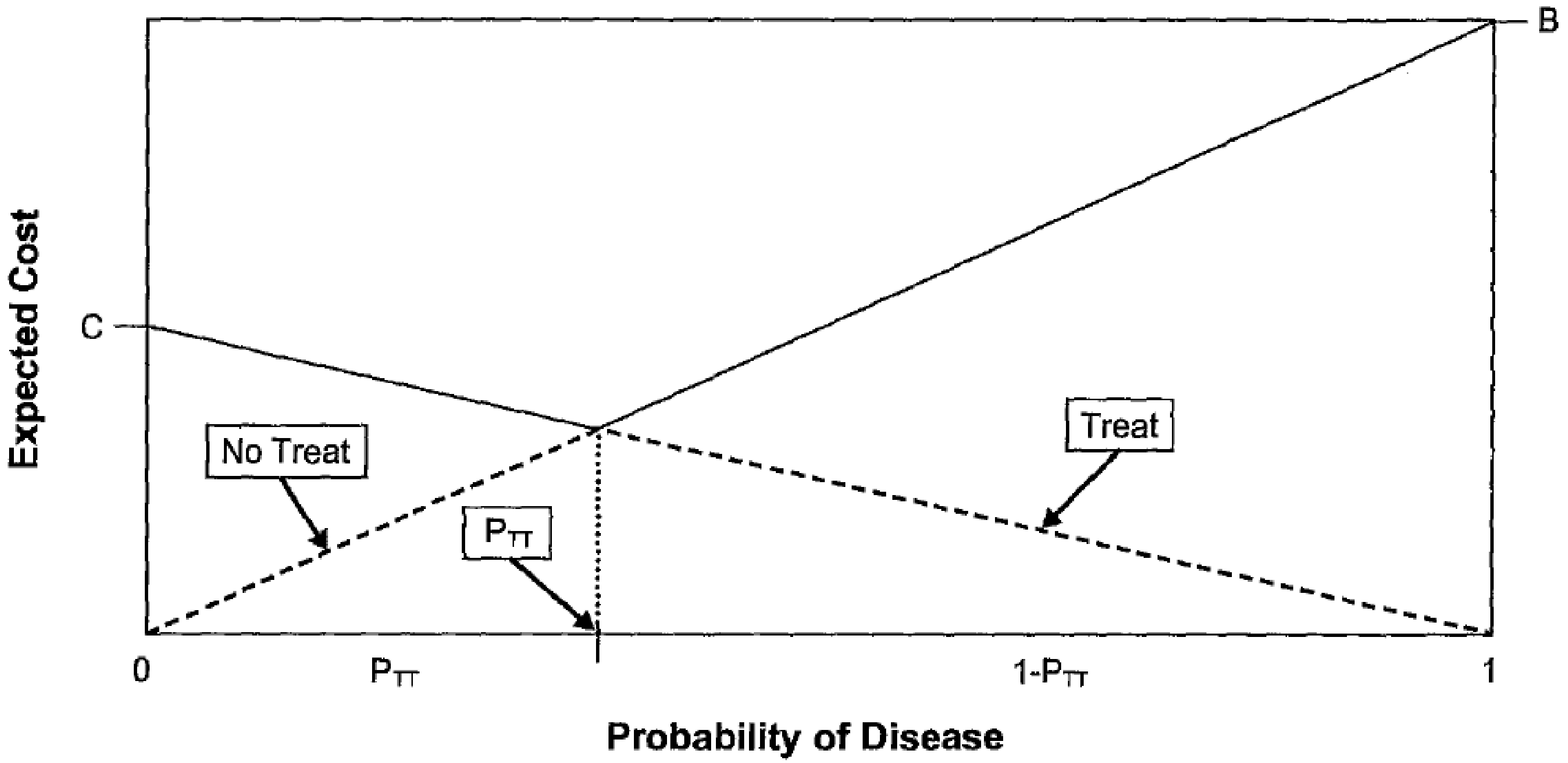


Figure 3.2 Expected costs of not treating and treating by probability of disease. For probabilities from 0 to P_{TT}, "No Treat" has the lowest expected cost. For probabilities from P_{TT} to 1, "Treat" has the lowest expected cost.

Treatment Threshold Probability (PTT)

- P_{TT} is the probability of disease at which:

$$P_{TT} \times B = (1 - P_{TT}) \times C$$

And therefore, the treatment threshold odds are given by:

$$\frac{P_{TT}}{(1 - P_{TT})} = \frac{C}{B}$$

and the threshold probability is

$$P_{TT} = \frac{C}{(C + B)}$$

- E.g. treating someone who does not have the disease is half as bad as failing to treat someone who does have the disease – should be willing to treat 2 people without disease to avoid failing to treat one person who has it
  - $C=1/2B$; $B=2xC$; $P_{TT} = C/(C + 2C) = C/3C= 1/3 = 0.33$

What Result Should Prompt Treatment?

Need to know the relative cost of errors: treating unnecessarily (C) versus failing to treat (B)

- Assume B = 4C

  Threshold Odds = c/b = c/4c = 0.25

  P_{TT} = c/(c+b) = c/(c+4c) = c/5c = 0.2

- Starting with P = 0.38

  Pretest Odds = 0.38/0.62 = 0.61

Optimal Cutoff = r*

- Newman and Kohn in *Evidence-Based Diagnosis (p. 82)* advocate setting the optimal cutoff r* as **the least abnormal** test result (r) such that

  Post-Test Odds (of disease) ≥ Treatment Threshold Odds

  Pre-Test Odds (of disease) × LR(r*) ≥ Treatment Threshold Odds (C/B)

  $[P/(1-P)] \times LR(r^*) \geq C/B$

- LR(r*) would need to be at least:

  Threshold Odds (C/B) divided by Pretest Odds

What Result Should Prompt Treatment?

- **P_{TT} = 0.2 → Threshold odds = 0.25; Pretest Odds =0.61**
- **LR(r) must be at least 0.25/0.61 = 0.41**

WBC (/uL) Interval	% of D+	% of D-	Interval LR	Post Test Prob
>100,000	29%	1%	29	0.95
50,001-100,000	33%	7%	4.7	0.74
25,001-50,000	15%	19%	0.8	0.33
0 - 25,000	23%	73%	0.3	0.16

TREAT

NO TREAT

What Result Should Prompt Treatment?

- **P_{TT} = 0.2 → Threshold odds = 0.25**
- **Pre-Test Probability = 0.04, not 0.38; Pretest Odds =0.042**
- **LR(r) must be at least 0.25/0.042 = 5.95**

WBC (/uL) Interval	% of D+	% of D-	Interval LR	Post Test Prob	
>100,000	29%	1%	29	0.55	TREAT ↑
50,001-100,000	33%	7%	4.7	0.16	NO TREAT
25,001-50,000	15%	19%	0.8	0.03	
0 - 25,000	23%	73%	0.3	0.01	

Sample Size Calculation

Sample Size Considerations: Confidence Interval for AUC

- Assume test results (or after some unknown monotonic transformation) follow a binormal distribution
  - i.e., separate normal distribution for diseased and non-diseased subjects

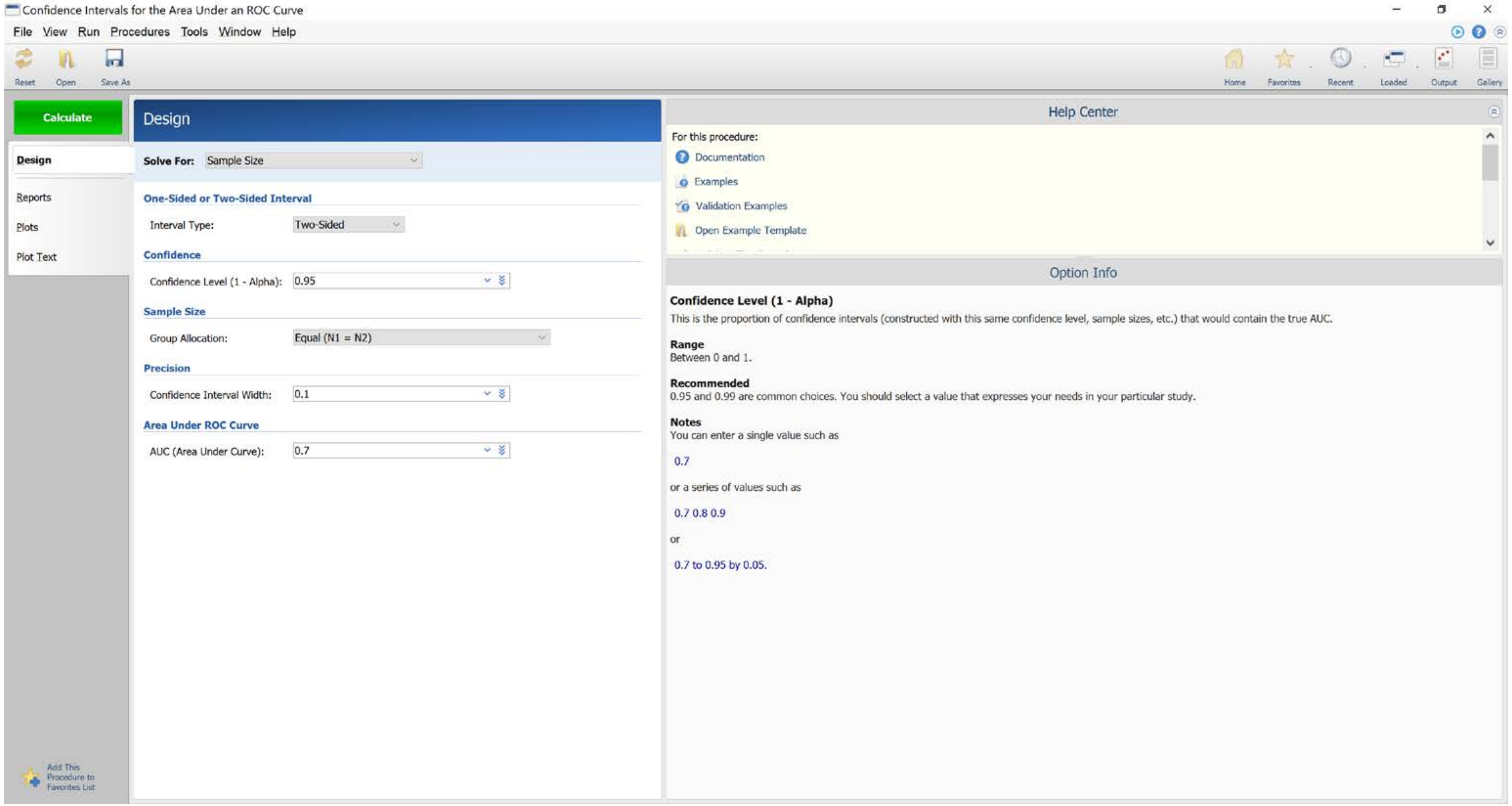
$$AUC = \int TPR(c)FPR'(c)dc$$

- Parameters and covariance matrix are estimated by maximum likelihood estimation
- Estimated AUC is asymptotically normal, and confidence interval is
$$AUC \pm z_{\alpha/2}SE(AUC)$$

Sample Size Considerations: Confidence Interval for AUC

- Example
    - Estimated AUC = 0.70
    - Two-sided, 95% confidence level
    - Confidence interval width = 0.10
    - # patients without disease = # patients with disease
    - What is the required sample size?

Sample Size Considerations: Confidence Interval for AUC (PASS input)



Sample Size Considerations: Confidence Interval for AUC (PASS output)

Confidence Intervals for the Area Under an ROC Curve

Numeric Results for Two-Sided Confidence Interval for ROC Curve's AUC

Confidence Level	Total Subjects N	Ratio N2/N1 R	Number Positive N1	Number Negative N2	Sample AUC	C.I. Width UCL-LCL	Lower Conf Limit LCL	Upper Conf Limit UCL
0.950	416	1.000	208	208	0.700	0.100	0.650	0.750

Report Definitions

Confidence Level is the proportion of confidence intervals (constructed with this same confidence level, sample size, etc.) that would contain the true coefficient alpha.
N is the total number of subjects sampled.
R is N2 / N1, so that N2 = R x N1.
N1 is the number of subjects sampled from the 'positive' group.
N2 is the number of subjects sampled from the 'negative' group.
Sample AUC is the anticipated value of the sample area under the ROC curve.
C.I. Width (UCL-LCL) is the width of the confidence interval. It is the distance from the lower limit to the upper limit.
Lower and Upper Confidence Limits are the actual limits that would result from a dataset with these statistics. They may not be exactly equal to the specified values because of the discrete nature of the N1 and N2.

References

Hanley, J.A. and NcNeil, B.J. 1982. 'The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve.' Radiology, Vol 148, 29-36.
Kryzanowski, W.J. and Hand, D.J. 2009. 'ROC Curves for Continuous Data.' Chapman & Hall/CRC Press.

Summary Statements

A random sample of 208 subjects from the positive population and 208 subjects from the negative population produce a two-sided 95.0% confidence interval with a width of 0.100 when the sample AUC is 0.700.

Sample Size Considerations: Test for One ROC Curve

- $H_0: AUC = \theta_0$ vs. $H_1: AUC \neq \theta_0$
  - θ_0 is 0.5 (non-informative test) or the AUC for a standard test
- <u>Continuous</u> test results
  - Binormal distribution
  - To achieve power $1 - \beta$ at $AUC = \theta_1$ (for the new test) with type I error rate α, required sample size in the diseased group is

$$N_+ = \frac{\left(z_{\alpha/2}\sqrt{V(\theta_0)} + z_\beta\sqrt{V(\theta_1)}\right)^2}{(\theta_1 - \theta_0)^2}$$

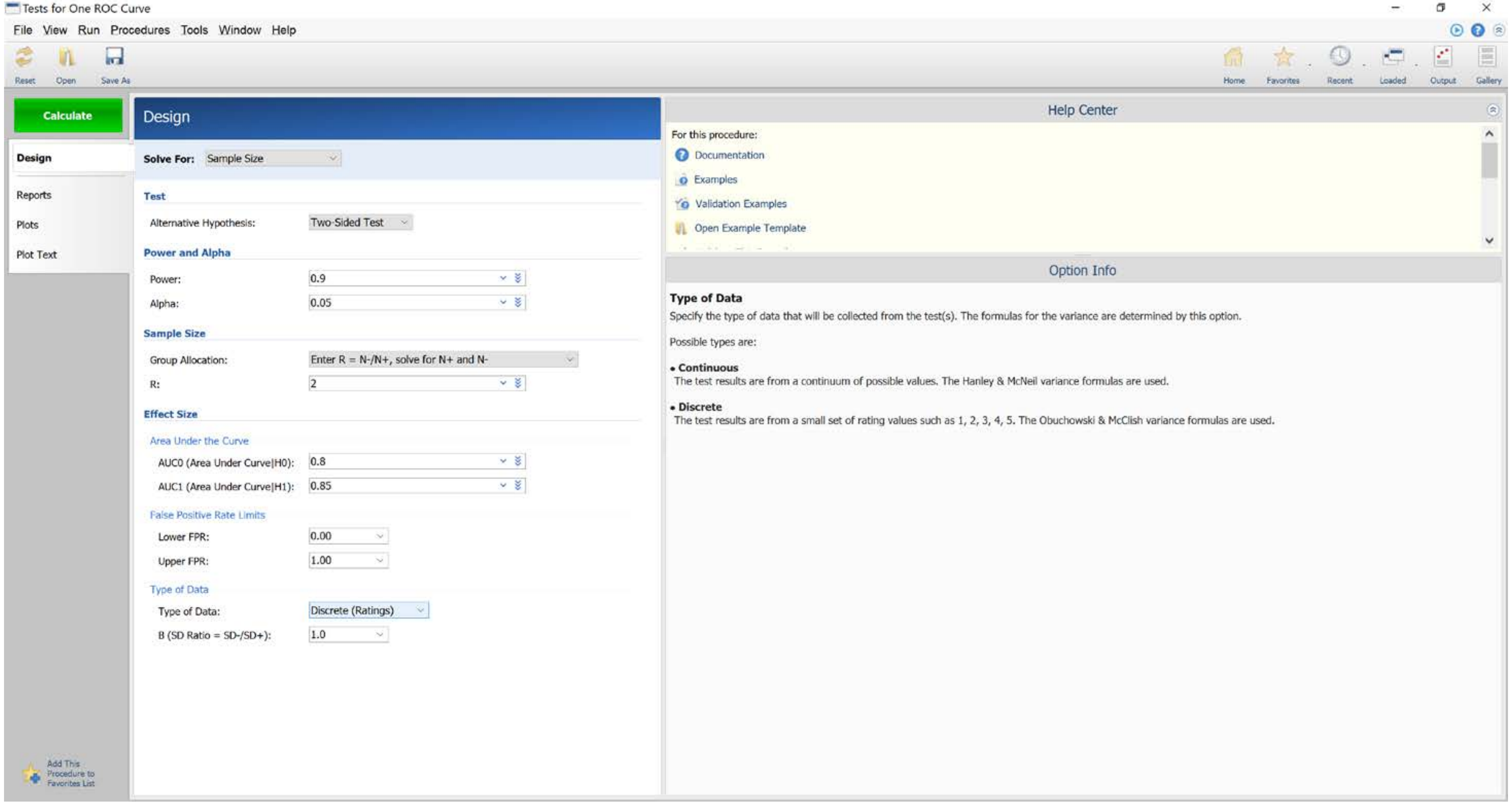
where V is the variance function of the estimated AUC.

- Sample size formula also available if test results are <u>discrete</u> ratings (Obuchowski, 1998)

Sample Size Considerations: Test for One ROC Curve

- Example
  - Test results measured on a discrete rating scale from 1 to 5
  - Standard test has AUC = 0.80
  - Wish to evaluate a new test with hypothesized AUC = 0.85
  - Two-sided test, type I error rate 0.05
  - 90% power
  - Patients without disease are twice as many as patients with disease
  - What is the required sample size?

Sample Size Considerations: Test for One ROC Curve (PASS input)



Sample Size Considerations: Test for One ROC Curve (PASS output)

Tests for One ROC Curve

Numeric Results for Testing AUC0 = AUC1 with Discrete (Rating) Data
Test Type = Two-Sided. FPR1 = 0.00. FPR2 = 1.00. B = 1.00.

Target Power	Actual Power	N+	N-	N	Target R	Actual R	AUC0'	AUC1'	Diff'	AUC0	AUC1	Diff	Alpha
0.90	0.90069	381	762	1143	2.00	2.00	0.8000	0.8500	0.0500	0.8000	0.8500	0.0500	0.050

References

Hanley, J. A. and McNeil, B. J. 1983. 'A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases.' Radiology, 148, 839-843. September, 1983.

Obuchowski, N. and McClish, D. 1997. 'Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices.' Statistics in Medicine, 16, pages 1529-1542.

Sample Size Considerations: Test for Two ROC Curves

- Compare AUC of two tests, obtained from the same patients
- Define $\Delta = \theta_1 - \theta_2$ to be the difference in AUC of the two tests
- $H_0: \Delta = 0$ vs $H_1: \Delta \neq 0$
- Let $\widehat{\Delta}$ be the maximum likelihood estimator of Δ
- To achieve power $1 - \beta$ at an alternative value Δ ($\neq 0$) with type I error rate α, required sample size in the diseased group is

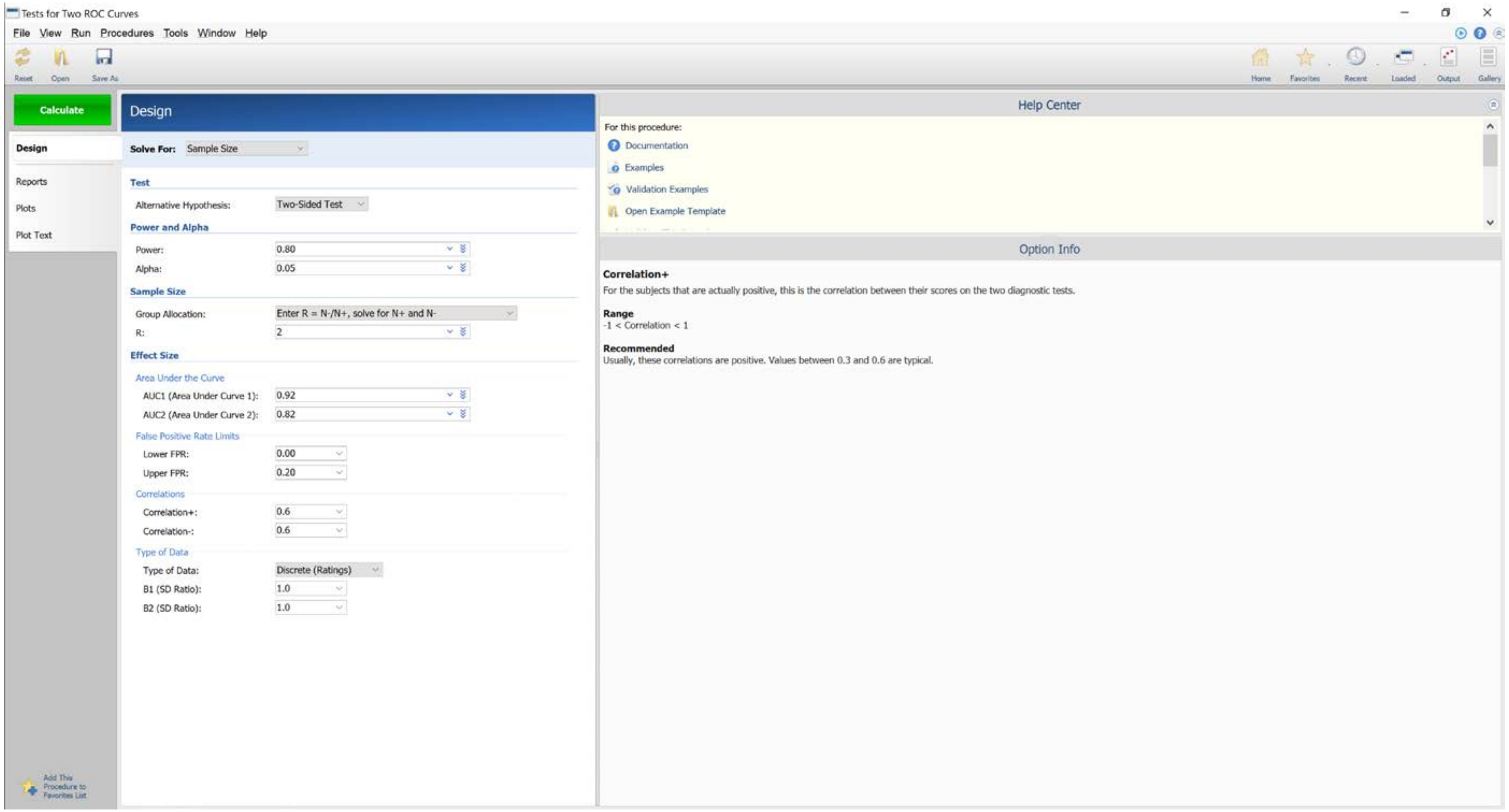
$$N_+ = \frac{\left(z_{\alpha/2}\sqrt{V_0\left(\widehat{\Delta}\right)} + z_\beta\sqrt{V_{Alt}\left(\widehat{\Delta}\right)}\right)^2}{\Delta^2}$$

- Different variance formulas for $V_0\left(\widehat{\Delta}\right)$ and $V_{Alt}\left(\widehat{\Delta}\right)$, depending on whether test results are continuous or discrete ratings

Sample Size Considerations: Test for Two ROC Curves

- Example taken from Obuchowski and McClish (1997)
- Compare automated classification system (AUC = 0.92) with an expert mammographer (AUC = 0.82) in finding malignant breast lesions
- Test results on discrete rating scale
- Restrict to FPR values from 0.0 to 0.2
- Patients without disease are twice as many as patients with disease
- Correlation between the two test results among diseased = correlation between the two test results among non-diseased = 0.6
- Two-sided test, type I error rate 0.05
- 80% power
- What is the required sample size?

Sample Size Considerations: Test for Two ROC Curves (PASS input)



Sample Size Considerations: Test for Two ROC Curves (PASS output)

Tests for Two ROC Curves

Numeric Results for Testing AUC1 = AUC2 with Discrete (Rating) Data

Test Type = Two-Sided. FPR1 = 0.0000. FPR2 = 0.2000. B1 = 1.0000. B2 = 1.0000. Corr+ = 0.6000. Corr- = 0.6000.

Target Power	Actual Power	N+	N-	N	Target R	Actual R	AUC1'	AUC2'	Diff'	AUC1	AUC2	Diff	Alpha
0.80	0.80080	117	234	351	2.00	2.00	0.9200	0.8200	-0.1000	0.1712	0.1352	-0.0360	0.050

References

Hanley, J. A. and McNeil, B. J. 1983. 'A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases.' Radiology, 148, 839-843. September, 1983.

Obuchowski, N. and McClish, D. 1997. 'Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices.' Statistics in Medicine, 16, pages 1529-1542.

Summary Statements

A sample of 117 from the positive group and 234 from the negative group achieve 80% power to detect a difference of 0.1000 between a diagnostic test with an area under the ROC curve (AUC) of 0.9200 and another diagnostic test with an AUC of 0.8200 using a two-sided z-test at a significance level of 0.050. The data are discrete (rating scale) responses. The AUC is computed between false positive rates of 0.00 and 0.20. The ratio of the standard deviation of the responses in the negative group to the standard deviation of the responses in the positive group for diagnostic test 1 is 1.00 and for diagnostic test 2 is 1.00. The correlation between the two diagnostic tests is assumed to be 0.60 for the positive group and 0.60 for the negative group.

Software Considerations

- For data analysis
  - STATA
  - SAS
    - Proc Logistic
    - Proc NLMixed
    - %ROC Macro
  - R (‘pROC’ package)
- For sample size calculation
  - PASS
  - R (‘pROC’ package)

ROC Analysis: Pros and Cons

- **Pros:**
- Provides a wholistic picture (a global assessment of a test’s accuracy)
- Not dependent on disease prevalence
- Does not force us to pick a single cut-off point
- Shows the trade off between Sn and Sp
- Great for comparing accuracy of competing tests
- Can be applied to any diagnostic/prognostic system

- **Cons:**
- Not very intuitive for clinicians; the ROC and AUC cannot be directly used for any given patient
- Clinicians prefer simple yes/no test results
- You can have the same AUC, but different shapes
- Does not fit into the EBM framework of working with LRs and probabilities
- Very hard to meta-analyze

Pai, McGill University

Acknowledgements

- Tabitha Garwe, PhD
- Sara Vesely, PhD
- Chao Xu, PhD
- Lance Ford, PhD

Acknowledgements

- This presentation includes content from "Epi 204: Clinical Epidemiology" by Drs. Michael Kohn and Tom Newman, accessed from https://epibiostat.ucsf.edu/clinical-epidemiology-epi-204

Selected References

1. Zweig, MH, Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 1993;39/4, 56-577.

2. DeLong, ER, DeLong, DM, Clarke-Pearson, DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44, 837-845.

3. Hanley, JA, McNeil, BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982, 143, 29-36.

4. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean J Radiol. 2004;5(1):11–18. doi:10.3348/kjr.2004.5.1.11

5. Newman, TB, Kohn, MA. Evidence-based Diagnosis: An Introduction to Clinical Epidemiology (2nd Edition). Cambridge University Press, Cambridge, UK, 125 (2020)

6. PASS 16 Power Analysis and Sample Size Software (2018). NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass.