Biostatistics for Health Care Researchers: A Short Course

Evaluation of Diagnostic Tests

Presented by:

Siu L. Hui, Ph.D.

Department of Medicine, Division of Biostatistics

Indiana University School of Medicine

Objectives

- Calculate and interpret sensitivity, specificity, predictive value positive, and predictive value negative.
- Understand the principle behind ROC curves.
- Understand the use of kappa and intraclass correlation coefficients to measure agreement.

Examples

- Screening
  - Lab (BMP, lipids)
  - Imaging (bone density, mammogram)
  - Questionnaires (depression, dementia, QOL)
- Diagnostic
  - Blood tests for infection
  - Imaging (X-ray, CT, MRI)
  - Histology

All tests have errors. How accurate are they?

Outline

- ***Accuracy*** of a diagnostic test:

  - Binary data: ***sensitivity*** and ***specificity***
  predictive value positive and negative

  - Ordinal or continuous data: ***ROC curve***

- Measure of the ***agreement*** of two tests:

  - nominal or ordinal data: ***Kappa***

  - Continuous data: ***Intraclass correlation coefficient***

Example 1: Staging of Prostate Cancer with MRI

Tempany, et al. (1994) studied the accuracy of conventional MRI in detecting advanced stage prostate cancer.

- Disease: advanced stage prostate cancer.
- Test: conventional MRI.
- The **true disease status** was established by surgery.

Question: How accurate is conventional MRI in detecting advanced stage prostate cancer?

Tempany, Zhou, Zerhouni, Rifkin, Quint, Piccoli, Ellis, and McNeil (1994). “Staging of Prostate Cancer: Results of Radiology Diagnostic Oncology Group Project Comparison of Three MR Imaging Techniques” Radiology, 192:47-54.

Example 1 (continued)

Disease\MRI	T+	T-	total
D+	70	45	115
D-	32	53	85
total	102	98	200

Accuracy of a Test

- **True Disease Status:**

  Disease (D^+) and non-disease (D^-) by **gold standard**

  {Advanced stage vs. early stage by surgery}

- **Test Result:**

  Positive (T^+) and negative (T^-) results from a test of interest.

  {Advanced stage vs. early stage as assessed by MRI}

Example 1 (continued)

Disease\MRI	T+	T-	total
D+	70	45	115
D-	32	53	85
total	102	98	200

Overall Accuracy

- N = total # of cases
- A = # of correctly diagnosed cases
- Overall accuracy = A/N

Example 1 (continued)

Disease\MRI	T+	T-	total
D+	70	45	115
D-	32	53	85
total	102	98	200

Overall Accuracy = (70 + 53)/200 = 0.615 ~ 62%

Sensitivity and Specificity

- ***Sensitivity*** (Sens) - the ability of a test to give a positive finding when the person tested truly has the disease under study.

- ***Specificity*** (Spec) - the ability of a test to give a negative finding when the person tested truly is free of the disease under study.

Sensitivity and Specificity

$$\text{Sens} = P(T^{+}|D^{+}) = \frac{\#\text{ of }T^{+}\text{ and }D^{+}}{\#\text{ of }D^{+}}$$

$$\text{Spec} = P(T^{-}|D^{-}) = \frac{\#\text{ of }T^{-}\text{ and }D^{-}}{\#\text{ of }D^{-}}$$

Sensitivity and Specificity

- ***Sensitivity***: True positive rate (TPR)
- ***Specificity***: True negative rate (1-FPR)

where FPR=false positive rate

Example 1 (continued)

Disease\MRI	T+	T-	total
D+	**70**	**45**	**115**
D-	32	53	85
total	102	98	200

Sens = 70/115 = 61%

Example 1 (continued)

Disease\MRI	T+	T-	total
D+	70	45	115
D-	**32**	**53**	**85**
total	102	98	200

Spec = 53/85 = 62%

- In summary, **Sensitivity** and **Specificity** are two **intrinsic properties** of a test.

BUT, clinical providers have to infer a patient's disease status from test results.

- *How well can a given test result of a patient predict the disease status of the patient?*
  {How likely a patient with positive MRI result actually has advanced stage prostate cancer?}

Predictive Values

- ***Predictive value positive (PV⁺)*** is the probability that a patient with a positive test result actually has the disease:

$$\mathrm{PV}^{+} = \frac{\text{\# of diseased patients with a positive test}}{\text{\# of patients with a positive test}}$$

- ***Predictive value negative (PV⁻)*** is the probability that a patient with a negative test does not have the disease:

$$\mathrm{PV}^{-} = \frac{\text{\# of non-diseased patients with a negative test}}{\text{\# of patients with a negative test}}$$

PV$^+$ and PV$^-$

- Both PV$^+$ and PV$^-$ depend on the sensitivity, the specificity and the disease prevalence (Prev).

$$\mathrm{PV}^{+} = \mathrm{P}\left(\mathrm{D}^{+} | \mathrm{T}^{+}\right) = \frac{\mathrm{Sens} \times \mathrm{Prev}}{\mathrm{Sens} \times \mathrm{Prev} + \left(1 - \mathrm{Spec}\right) \times \left(1 - \mathrm{Prev}\right)}$$

$$\mathrm{PV}^{-} = \mathrm{P}\left(\mathrm{D}^{-} | \mathrm{T}^{-}\right) = \frac{\mathrm{Spec} \times \mathrm{Prev}}{\mathrm{Spec} \times \mathrm{Prev} + \left(1 - \mathrm{Sens}\right) \times \left(1 - \mathrm{Prev}\right)}$$

Example 2: A Screening Test for a Rare Disease

Disease\Test	T+	T-	**total**
D+	19	1	**20**
D-	99	1881	**1980**
total	118	1882	**2000**

- Disease prevalence=20/2000=1%.

Example 2: A Screening Test for a Rare Disease

Disease\Test	T+	T-	total
D+	19	1	20
D-	99	1881	1980
total	118	1882	2000

- Disease prevalence=20/2000=1%.
- Sens=19/20=95%
- Spec=1881/1980=95%.

Example 2: A Screening Test for a Rare Disease

Disease\Test	**T+**	T-	total
D+	**19**	1	20
D-	**99**	1881	1980
total	**118**	1882	2000

- Disease prevalence=20/2000=1%.
- Sens=19/20=95%, Spec=1881/1980=95%.
- PV^+=19/118=16.1% .

Example 2: A Screening Test for a Rare Disease

Disease\Test	T+	**T-**	total
D+	19	**1**	20
D-	99	**1881**	1980
total	118	**1882**	2000

- Disease prevalence=20/2000=1%.
- Sens=19/20=95%, Spec=1881/1980=95%.
- PV^+=19/118=16.1%, PV^-=1881/1882=99.9%.

Example 2 (continued)

Prev(%)	PV^+ (%)	PV^- (%)
1	16.1	99.9
5	50.0	99.7
20	82.6	98.7
50	95.0	95.0
75	98.3	83.7

- Sens = Spec = 95%.
- Note how the prevalence affects PV^+ and PV^-.
- *When does it not make sense to screen?*

- *What if your test gives a continuous reading rather than +/- ?*

Example 3: Blood Test for Disease

ID	D	T	≤1	≤2	≤3	≤4
1	-	0.2	-	-	-	-
2	-	0.7	-	-	-	-
3	-	1.8	+	-	-	-
4	-	2.0	+	-	-	-
5	-	3.1	+	+	-	-
6	-	3.3	+	+	+	-
7	+	1.5	+	-	-	-
8	+	2.4	+	+	-	-
9	+	3.0	+	+	+	-
10	+	3.1	+	+	+	-
11	+	3.8	+	+	+	-
12	+	4.0	+	+	+	-
Sens			1.00	0.83	0.67	0.00
Spec			0.33	0.67	0.83	1.00
FPR			0.67	0.33	0.17	0.00

ROC Curve

(Receiver Operating Characteristic Curve)

- *When is it applied?*
  - Test results are ***continuous*** and we may have more than one possible cutoff points, or
  - We have multiple degrees of suspicion for a given a test (***ordinal*** response, e.g. definitely no disease, probably no disease, probably disease, and definitely disease).
- *How is it plotted?* FPR on the horizontal axis and TPR on the vertical axis.

Recall: True positive rate (TPR) = Sens

False positive rate (FPR) = 1-Spec

ROC Curve

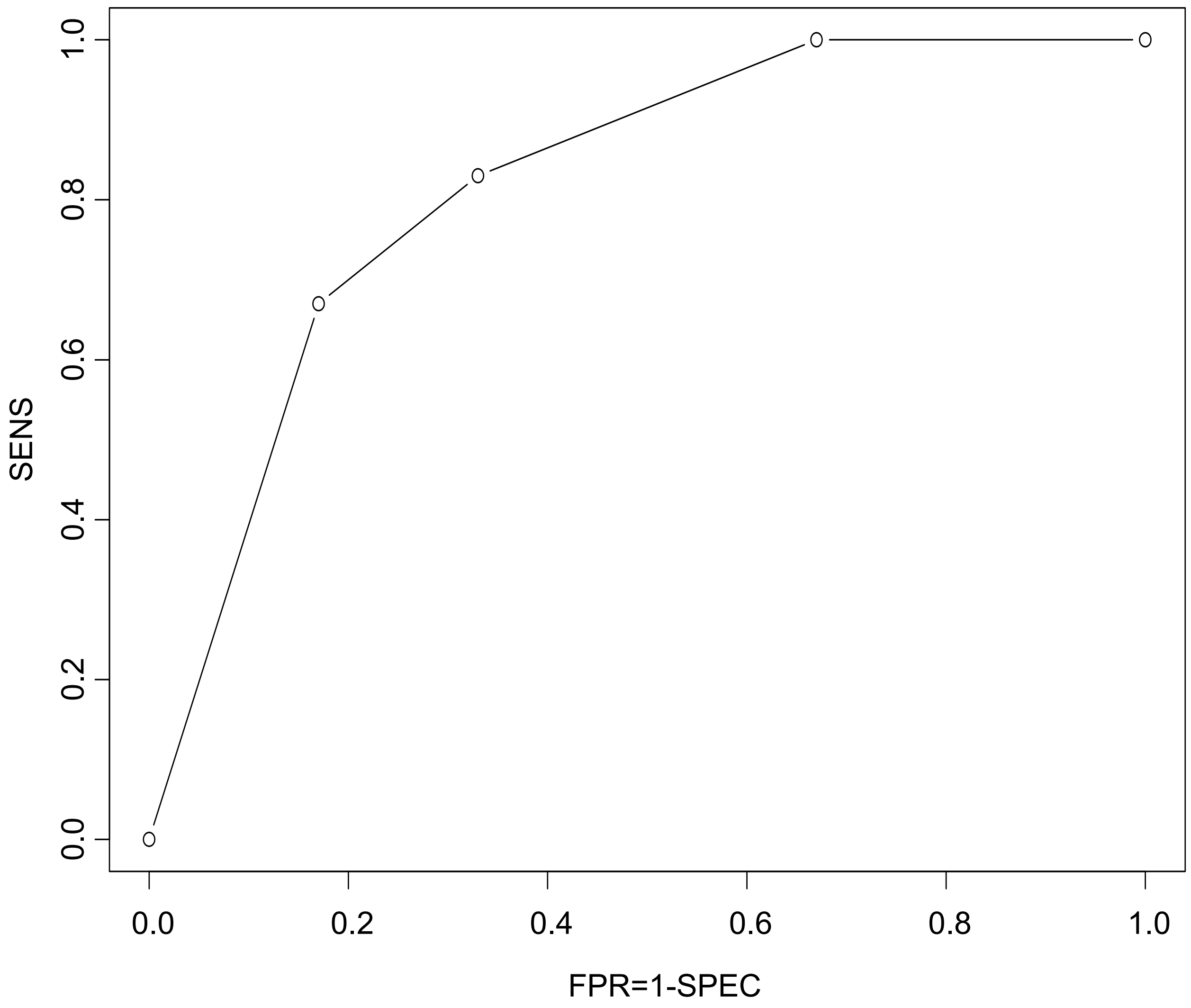
(Receiver Operating Characteristic Curve)

- *When is it applied?*
  - Test results are ***continuous*** and we may have more than one possible cutoff points, or
  - We have multiple degrees of suspicion for a given a test (***ordinal*** response, e.g. definitely no disease, probably no disease, probably disease, and definitely disease).
- *How is it plotted?* FPR on the horizontal axis and TPR on the vertical axis.

Recall: True positive rate (TPR) = Sens

False positive rate (FPR) = 1-Spec

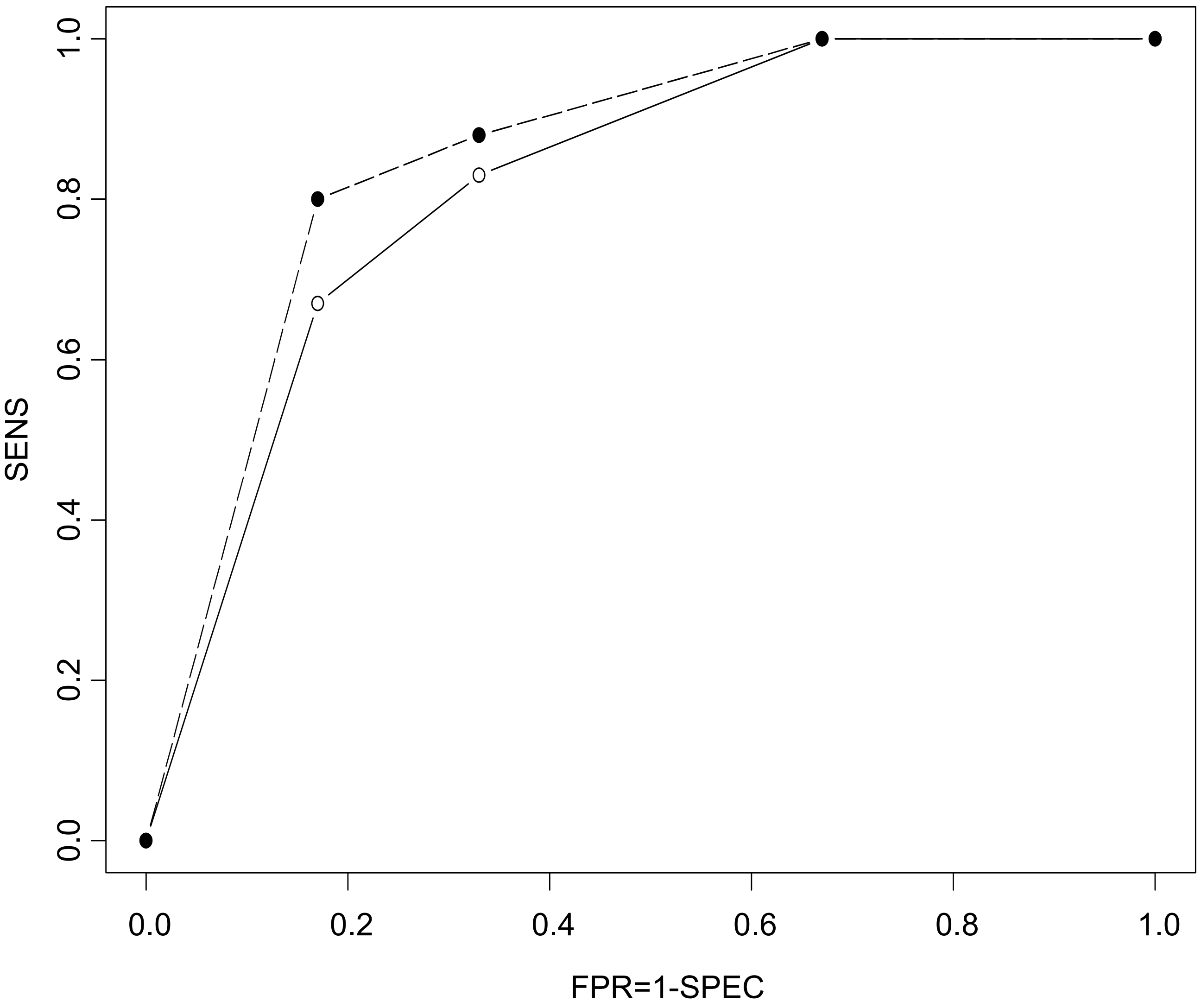
Example 3 (continued)



ROC Curves

- *Why do we need the ROC curve?*
  - Displays Sens (benefit) and FPR (cost) under different thresholds so decision makers can choose the appropriate threshold for their situation.
  - *How to choose a threshold in practice?* Based on how the test is used e.g. screening vs. diagnostic
  - Provides the ability to compare two or more diagnostic tests.

ROC Curve Comparison



Comments

- We can compare the accuracies of two tests if we know the gold standard.

- *What if we don’t have the gold standard?*

Reliability

- When the gold standard is not available, a test is considered reliable if it agrees with another **reference** test.
- A test is considered reliable if it provides higher inter-rater and intra-rater (or inter-assay and intra-assay) agreement.
- ***Kappa***: used for nominal or ordinal data agreement
- ***ICC***: used for continuous data agreement

Example 4: Biphasic Radiography vs. Fiberoptic Endoscopy in Gastric Ulcers

Endo\Radio	No Ulcer	Ulcer	total
No Ulcer	351	4	355
Ulcer	7	12	19
total	358	16	374

Shaw, van Romunde, Griffioen, Janssens, Kreuning, Eilers (1987). “Peptic Ulcers and Gastric Carcinoma: Diagnosis with Biphasic Radiography Compared with Fiberoptic Endoscopy” Radiology, 163:39-42.

Kappa (κ)

- P_o=observed agreement.
- P_e=agreement expected by chance.
- P_o-P_e=agreement beyond chance.
- 1-P_e=the maximum agreement possible beyond chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

An Interpretation of Kappa

Kappa	Strength of Agreement
<0.00	poor
0.00-0.20	Slightly poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Example 4 (continued)

Endo\Radio	No ulcer	Ulcer	total
No Ulcer	351	4	355
Ulcer	7	12	19
total	358	16	374

- Total observed agreement = (351+12)/374 = 97%.
- κ=.67 (Substantial).

Intraclass Correlation

- ICC compares the variability of a trait between subjects to the total variation across all ratings and all subjects.

- As the variability of a trait between subjects increases relative to the total variability, ICC moves closer to 1.

Example 5: Faculty Ratings of Residents' Performances

Resident	Faculty1	Faculty2
1	77	81
2	80	79
3	55	60
4	91	88
5	60	62
6	80	84

ICC=.96

Summary

- Measure of ***accuracy*** with gold standard:
  - Sens, Spec, PV+, PV-
  - ROC curve

- Measure the ***agreement*** of two tests in the absence of gold standard:
  - Kappa
  - ICC