NUFFIELD DEPARTMENT OF
PRIMARY CARE
HEALTH SCIENCES

UNIVERSITY OF
OXFORD

# Statistics for Diagnostic Accuracy

## Diagnostic Evidence Workshop

Jason Oke

2/10/2015

Why should a doctor need to know how to calculate the chance of breast cancer in a patient with a positive mammogram, given "a prevalence of 1%, a *sensitivity* of 90%, and a *false positive rate* of 9%"? ……

What the doctor needs is a test that gives a straight yes or no answer, or something close to it.

Shuster, S. BMJ 2011, 342:d2579: *The real problem is the biomedical ignorance of statisticians*
*http://www.bmj.com/content/342/bmj.d2579/rapid-responses*

# Motivation - research

- Diagnostic tests are important and costly
- Methodological quality of studies *was* and can still be poor.
    - Not generalisable
    - Badly designed – biased
    - Small samples - uncertain results

# Motivation – putting into practice

Results can be difficult to interpret

- Numerous and confusing terminology
- Interpreting results involves *statistical thinking*
- Often not clear how to balance trade-offs
- Value often depends on context – not just intrinsic accuracy .

# Basics

- Intrinsic accuracy of an *index* test is defined by how often it agrees with the *reference test* result
- Two results for each patient studied:
  - Reference test result - true status as determined by a *gold standard* test!
  - Index test result

Jargon counter - 5

# Basic 2 x 2 table

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | | |
| Negative | | | |
| Total | | | |

# Top left corner

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | | |
| Negative | | | |
| Total | | | |

# True positives

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** | | |
| Negative | | | |
| Total | | | |

# False Negatives

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | | |
| Negative | **FN** | | |
| Total | | | |

# TP + FP = Number with disease/condition present

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** | | |
| Negative | **FN** | | |
| Total | **TP + FN = No D+** | | |

# True negatives

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | **Absent** | Total |
| Positive | | | |
| Negative | | **TN** | |
| Total | | | |

# False positives

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | **Absent** | Total |
| Positive | | **FP** | |
| Negative | | | |
| Total | | | |

# TN + FP = Number without disease/condition absent

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | **Absent** | Total |
| Positive | | **FP** | |
| Negative | | **TN** | |
| Total | | **TN + FP = No D-** | |

# The good diagonal

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** | **FP** | |
| Negative | **FN** | **TN** | |
| Total | | | **N** |

# The bad diagonal

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** | **FP** | |
| Negative | **FN** | **TN** | |
| Total | | | **N** |

# Exercise 1: Binary tests

- Scenario : A study looked at the accuracy of MRI to detect thoracic aortic dissection. All 114 patients were assessed using an appropriate gold standard test and using MRI. One reader interpreted the MRI, without knowledge of the true status

- Reference status - 45 patients had a dissection, 69 did not.
- MRI results were categorised as;

1. Definitely not
2. Probably not
3. Possible dissection
4. Probable dissection
5. Definite dissection

| Dissection status (Reference test) | MRI result | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Present | 7 | 7 | 3 | 5 | 23 |
| Absent | 39 | 19 | 9 | 1 | 1 |

Using a cutpoint for the index test of

- MRI < 5 - Negative result
- MRI >= 5 - Positive result

Construct the 2x2 table of counts.

| Dissection status (Reference test) | MRI result | | | | |
|---|---|---|---|---|---|
| | Negative | | | | Positive |
| Present | 7 | 7 | 3 | 5 | 23 |
| Absent | 39 | 19 | 9 | 1 | 1 |

Task: Using a cutpoint for the index test of

- MRI < 5 - Negative result
- MRI >= 5 - Positive result

Construct the 2x2 table of counts.

| Dissection status (Reference test) | MRI result | | | | |
|---|---|---|---|---|---|
| | Negative | | | | Positive |
| Present | 7 | 7 | 3 | 5 | 23 |
| Absent | 39 | 19 | 9 | 1 | 1 |

Task: Using a cutpoint for the index test of

- MRI < 5 - Negative result
- MRI >= 5 - Positive result

Construct the 2x2 table of counts.

| Dissection status (Reference test) | MRI result | |
|---|---|---|
| | Negative | Positive |
| Present | 7 + 7 + 3 + 5 | 23 |
| Absent | 39 + 19 + 9 + 1 | 1 |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | | |
| Negative | | | |
| Total | | | |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** = 23 | | |
| Negative | | | |
| Total | | | |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** = 23 | | |
| Negative | **FN** = 22 | | |
| Total | | | |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** = 23 | | |
| Negative | **FN** = 22 | | |
| Total | **TP + FN** = 45 | | |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | | |
| Negative | | **TN = 68** | |
| Total | | | |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | **FP = 1** | |
| Negative | | **TN = 68** | |
| Total | | **FP + TN = 69** | |

# MRI for assessing aortic dissection

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP = 23** | **FP = 1** | T+ = 24 |
| Negative | **FN = 22** | **TN = 68** | T- = 90 |
| Total | **TP + FN = 45** | **FP + TN = 69** | **TP + FN + FP + TN = 114** |

# Measures of diagnostic accuracy

- Sensitivity
  - Refers to the diseased population
  - The ability of the index test to detect the condition/disease.
  - Also known as the true positive rate (TPR)
- Specificity
  - Refers to the un-diseased population
  - The index tests ability to exclude the condition
  - The same as the true negative rate (TNR)

- False positive rate (FPR) = 1 – specificity
- False negative rate (FNR) = 1 – sensitivity
- Accuracy – Probability of a correct result
- Youden's index – likelihood of a positive result among patients with versus without the condition.
- Diagnostic Odds Ratio – ratio of risk of a positive test amongst diseased patients and the risk of a positive test in the un-diseased patients

# Exercise:

## Calculate sensitivity and specificity:

Sensitivity = True positives / (True positives + False negatives)

- TP / (TP + FN) or TP / D+

Specificity = True negatives / (True negatives + False positives)

- TN / (TN + FP) or TN / D-

Accuracy = (True negatives + True Positives) / all tests

- (TN + TP) / N

Youden's index = Sensitivity + Specificity - 1

| MRI Test Result | Dissection status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | TP = 23 | FP = 1 | T+ = 24 |
| Negative | FN = 22 | TN = 68 | T- = 90 |
| Total | D+ = 45 | D- = 69 | N = 114 |

Sensitivity: 23/45 = 0.51

Specificity: 68/69 = 0.99

FPR : 1 – 0.99 = 0.01

FNR = 1 – 0.51 = 0.49

Accuracy = (23 + 68)/114 = 0.80

Youden's index = 0.51 + 0.99 – 1 = 0.5

# Discussion

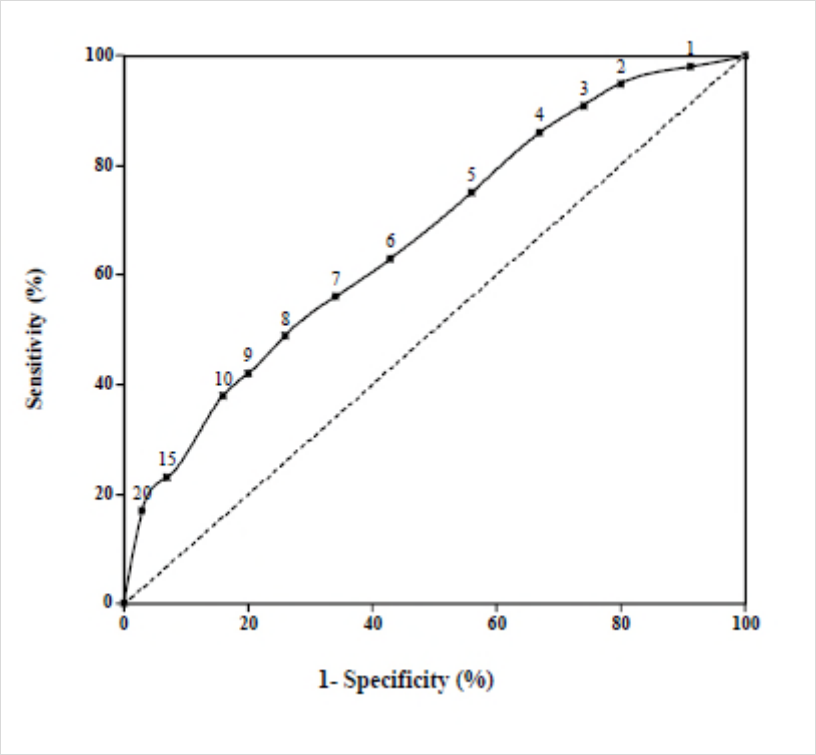
- Based on intrinsic accuracy (the sensitivity, specificity, FPR and FNR) do you think MRI for detection of aortic dissection have any value as diagnostic test?
- If were moved the MRI threshold from 5 to 4, what do you think would have happened?

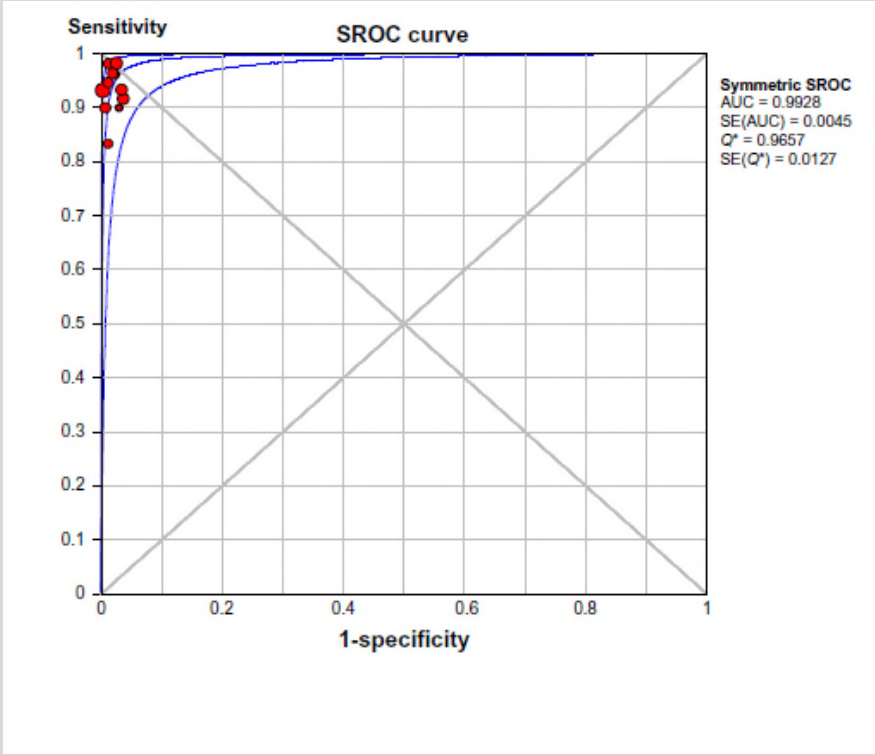
# The Receiver Operating Characteristic curve

- Many diagnostic results yield a numeric measurement (or ordinal scale) rather than +/-
- In the aortic dissection example we arbitrarily choose a decision threshold to define +/-.
- Sensitivity and specificity was affected by our choice of decision threshold.
- As specificity decreased the sensitivity increased – they are inherently linked.

# The ROC curve

- ROC curve method overcomes the limitations of a single sensitivity and specificity pair or summary measure.
- The curve is a plot of sensitivity (y axis) versus FPR (x  axis)
- Each point on the graph represents a different decision threshold.




Prostate-specific antigen (PSA)(widely used)

Non-Radiologist performed ultrasound for AAA (not used to screen women)

# Exercise 2: ROC curve

- Scenario : Consider a digital-imaging algorithm to identify patients whose implanted artificial heart valves have fractured.
- One measure to distinguish fractured valves is to determine the width of the gap between the valve strut legs using digital imaging. Larger gaps are associated with higher chance of fracturing.

# Exercise 2: ROC curve

- Twenty patients underwent elective surgery for valve replacement. 10 patients were found to have fractured valves, 10 did not.

| Fractured | Intact |
|---|---|
| 0.58 | 0.13 |
| 0.41 | 0.13 |
| 0.18 | 0.07 |
| 0.15 | 0.05 |
| 0.15 | 0.03 |
| 0.10 | 0.03 |
| 0.07 | 0.03 |
| 0.07 | 0.00 |
| 0.05 | 0.00 |
| 0.03 | 0.00 |

Gap measurements 10 patients with fractured heart valves and 10 patients without fractured valves.

# Exercise 2: Calculate the empirical ROC curve

1. Construct a 2x2 table for each unique value in the data.
2. Drive the sensitivity and FPR at each point
3. Plot these pairs on the graph.

Start with lowest possible threshold
(Test is positive if gap >= 0.00)

| Test = Negative | Test = Positive | |
|---|---|---|
| . | 0.13 | 0.58 |
| | 0.13 | 0.41 |
| | 0.07 | 0.18 |
| | 0.05 | 0.15 |
| | 0.03 | 0.15 |
| | 0.03 | 0.10 |
| | 0.03 | 0.07 |
| | 0.00 | 0.07 |
| | 0.00 | 0.05 |
| | 0.00 | 0.03 |

Green indicates not fractured, red - fractured

# Decision threshold test is positive if gap >= 0.00

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| Positive (>0.00) | TP = 10 | FN = 10 | 20 |
| Negative (<0.00) | FN = 0 | TN = 0 | 0 |
| Total | 10 | 10 | N = 20 |

Sensitivity: 10/10 = 1

FPR : 10/10 = 1

Threshold = 0.03

| Test = Negative | Test = Positive | |
|---|---|---|
| 0.00 | 0.13 | 0.58 |
| 0.00 | 0.13 | 0.41 |
| 0.00 | 0.07 | 0.18 |
| | 0.05 | 0.15 |
| | 0.03 | 0.15 |
| | 0.03 | 0.10 |
| | 0.03 | 0.07 |
| | | 0.07 |
| | | 0.05 |
| | | 0.03 |

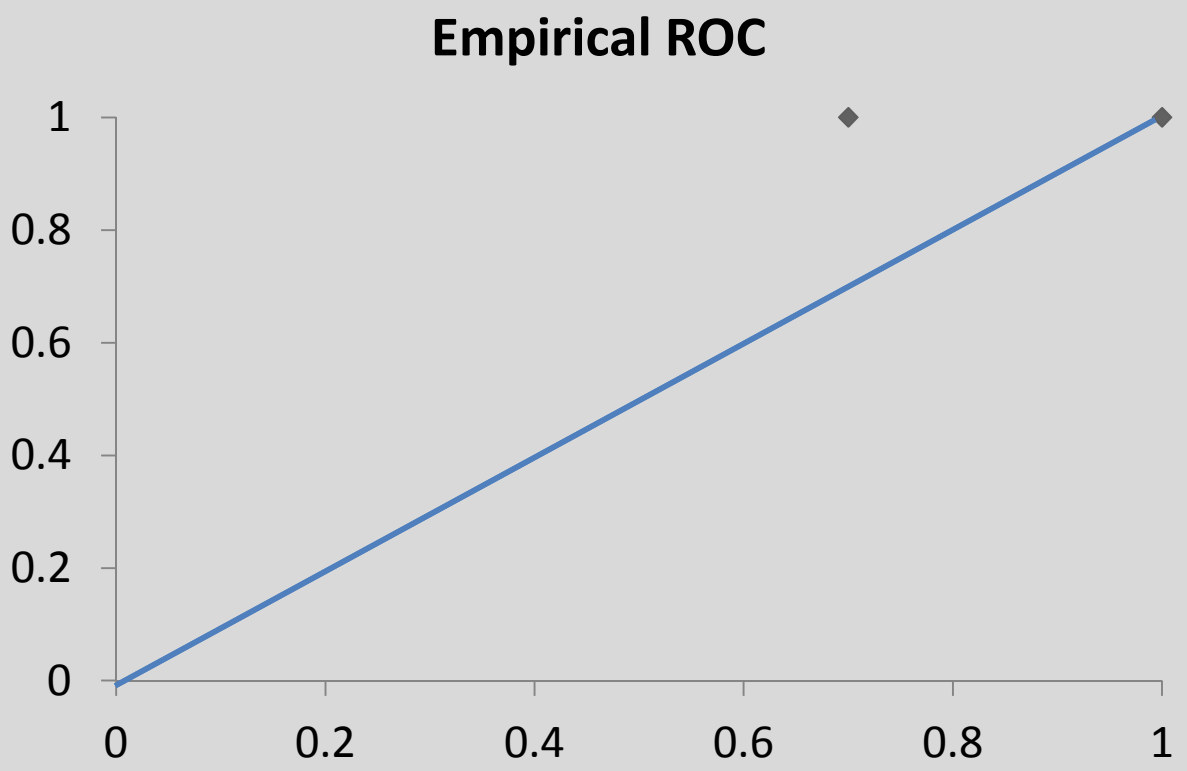
Threshold of 0.03

# Decision threshold test is positive if gap >= 0.03

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.03 | TP = 10 | FP = 7 | 17 |
| < 0.03 | FN = 0 | TN = 3 | 3 |
| Total | 10 | 10 | N = 20 |

Sensitivity: 10/10 = 1

FPR : 7/10 = 0.7


Empirical ROC
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

Threshold = 0.05

| Test = Negative | Test = Positive | |
|---|---|---|
| 0.00 | 0.13 | 0.58 |
| 0.00 | 0.13 | 0.41 |
| 0.00 | 0.07 | 0.18 |
| 0.03 | 0.05 | 0.15 |
| 0.03 | | 0.15 |
| 0.03 | | 0.10 |
| 0.03 | | 0.07 |
| | | 0.07 |
| | | 0.05 |
| | | |

Threshold of 0.05

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | | | |
| < 0.05 | | | |
| Total | 10 | 10 | N = 20 |

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | TP = 9 | | |
| < 0.05 | | | |
| Total | 10 | 10 | N = 20 |

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | TP = 9 | | |
| < 0.05 | FN = 1 | | |
| Total | 10 | 10 | N = 20 |

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | TP = 9 | | |
| < 0.05 | FN = 1 | TN = 6 | 7 |
| Total | 10 | 10 | N = 20 |

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | TP = 9 | FP = 4 | 13 |
| < 0.05 | FN = 1 | TN = 6 | 7 |
| Total | 10 | 10 | N = 20 |

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | TP = 9 | FP = 4 | 13 |
| < 0.05 | FN = 1 | TN = 6 | 7 |
| Total | 10 | 10 | N = 20 |

Sensitivity: 9/10 = 0.9

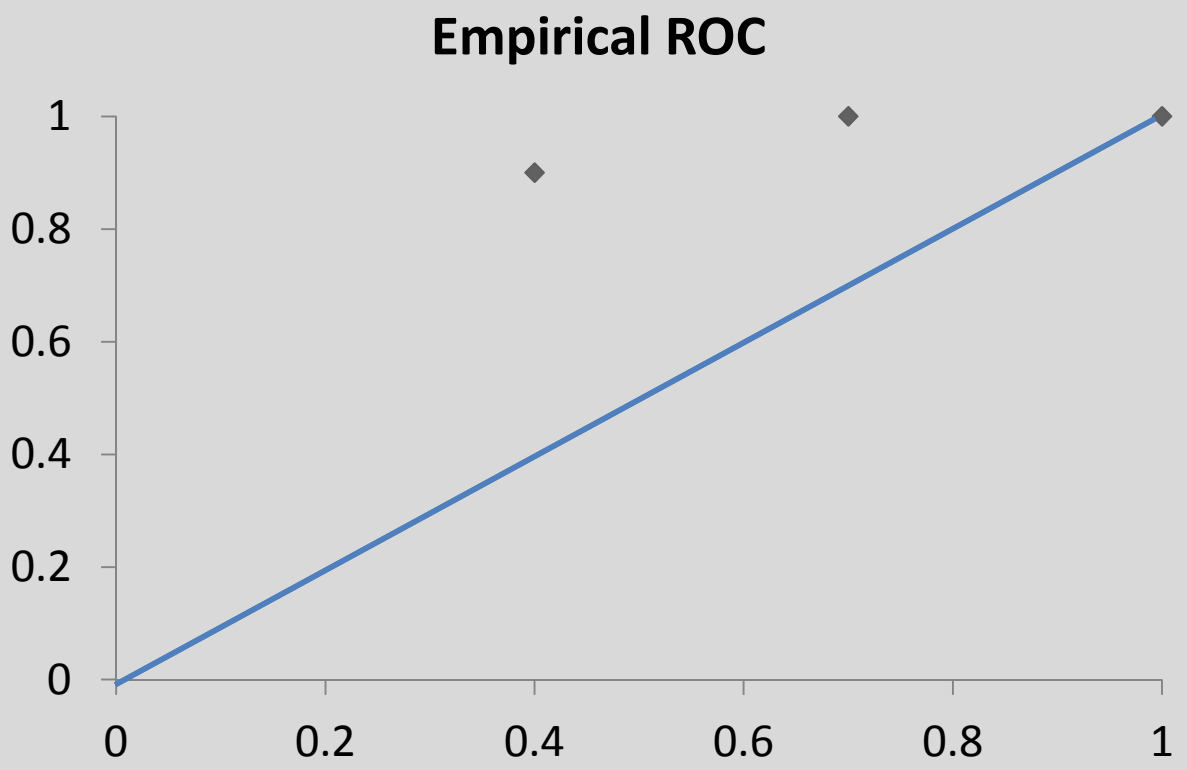
FPR : 4/10 = 0.4

# Decision threshold test is positive if gap >= 0.05

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.05 | TP = 9 | FP = 4 | 13 |
| < 0.05 | FN = 1 | TN = 6 | 7 |
| Total | 10 | 10 | N = 20 |

Sensitivity: 9/10 = 0.9

FPR : 4/10 = 0.4


Empirical ROC
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

| Test = Negative | Test = Positive | |
|---|---|---|
| 0.00 | 0.13 | 0.58 |
| 0.00 | 0.13 | 0.41 |
| 0.00 | 0.07 | 0.18 |
| 0.03 | | 0.15 |
| 0.03 | | 0.15 |
| 0.03 | | 0.10 |
| 0.03 | | 0.07 |
| 0.05 | | 0.07 |
| 0.05 | | |
| | | |

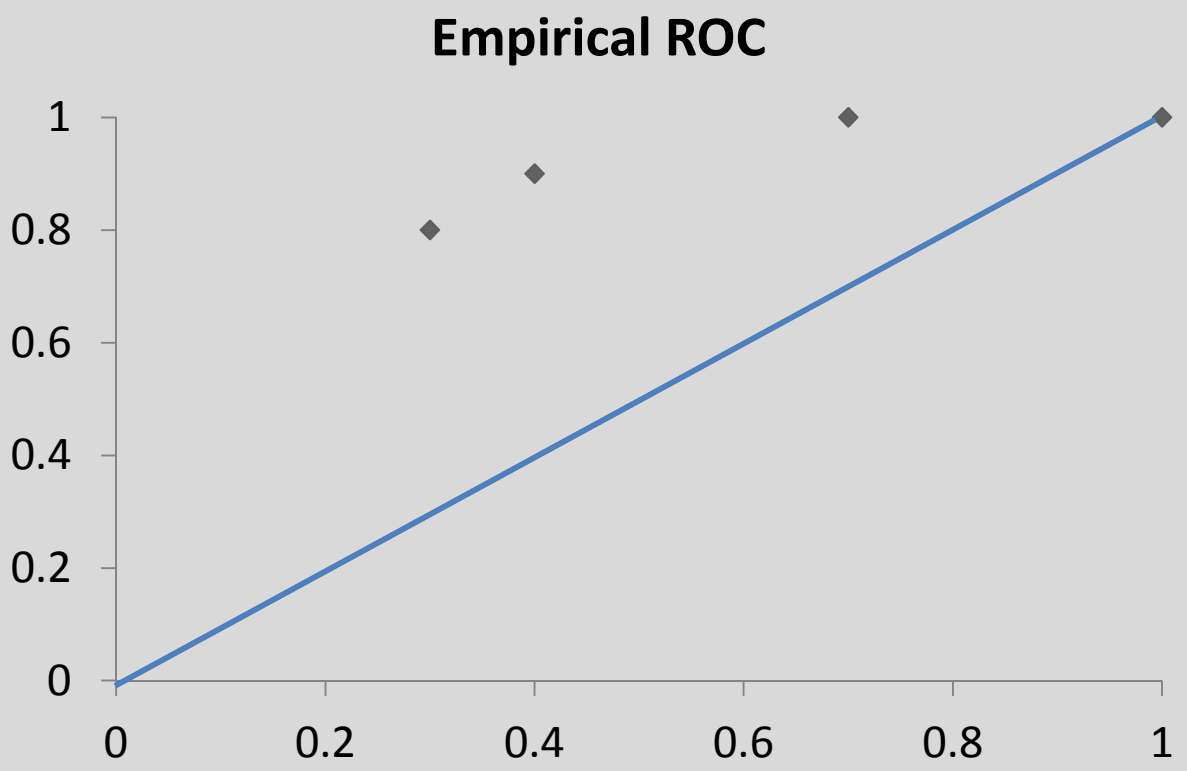
Threshold of 0.07

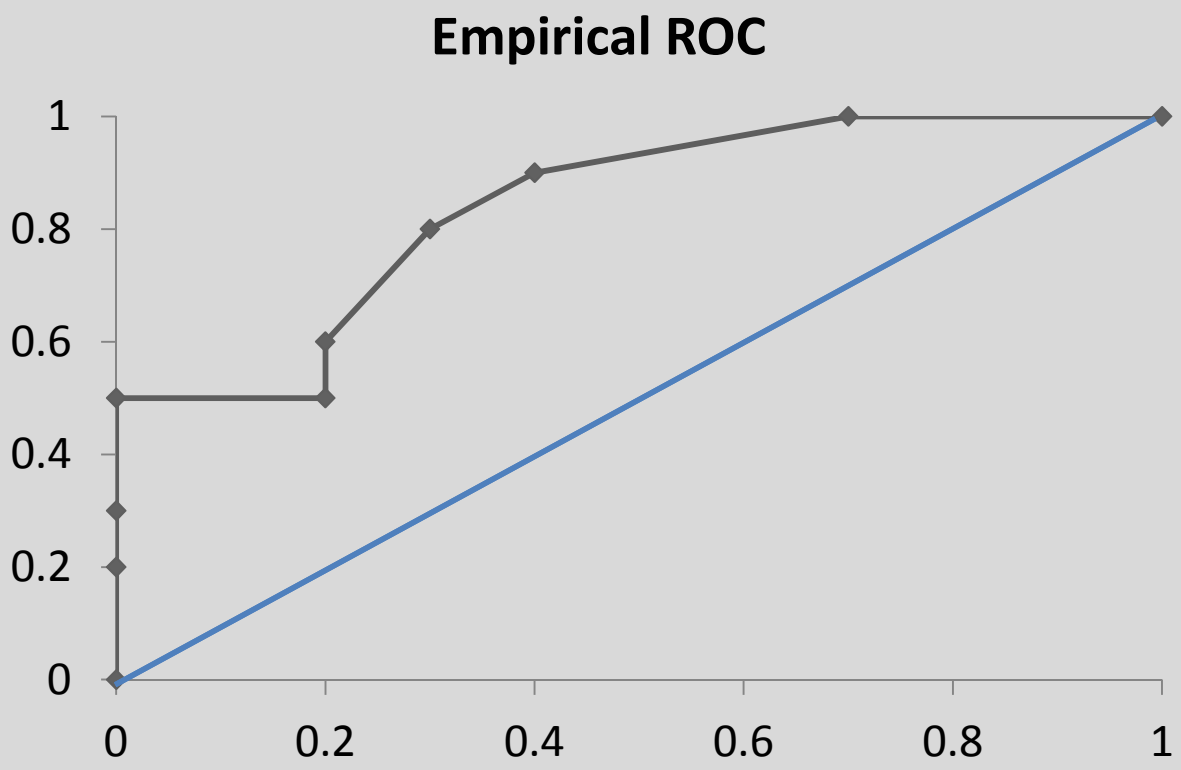
# Decision threshold test is positive if gap >= 0.07

| Gap | Status of heart valve | | |
|---|---|---|---|
| | Fractured | Intact | Total |
| >= 0.07 | TP = 8 | FP = 3 | 11 |
| < 0.07 | FN = 2 | TN = 7 | 9 |
| Total | 10 | 10 | N = 20 |

Sensitivity: 8/10 = 0.8

FPR : 3/10 = 0.3


Empirical ROC
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


Empirical ROC
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

# Why use a ROC curve

- A ROC curve is a visual representation of the data
- Does not require selection of a particular decision threshold
- Does not depend on the scale of the measurement, or the prevalence
- Very useful for comparing two or more similar tests

# Area under the ROC curve

The AUC is a single summary measure of the ROC Curve:

- Takes values between 0 and 1, but usually > 0.5
- Can be interpreted as;
  - Average Se for all possible values of Sp
  - Average Sp for all possible values of Se
  - The probability that a patient with the condition has a test result indicating greater suspicion than a patient without the condition

# Likelihood ratios (positive and negative)

## Positive likelihood ratio (LR +)

- How much more likely is a positive test result to be found in a person with the disease than in a person without it?

## Negative likelihood ratio (LR –)

- How much more likely is a negative test result to be found in a person without the disease than in a person with it?

# Calculation

Positive likelihood ration (LR +)

- TPR / FPR

Negative likelihood ratio (LR -)

- FNR / TNR

| Decision threshold | TPR | FPR | LR + |
|---|---|---|---|
| 0.00 | 1.00 | 1.00 | 1 |
| 0.03 | 1.00 | 0.7 | 1.43 |
| 0.05 | 0.9 | 0.4 | 2.25 |
| 0.07 | 0.8 | 0.3 | 2.67 |
| 0.1 | 0.6 | 0.2 | 3 |
| 0.13 | 0.5 | 0.2 | 2.5 |
| 0.15 | 0.5 | 0 | Undefined |

# Interpretation

Question: If the LR+ for film-screen mammography is 1.53. Does this imply that given a positive mammogram, a women is more 1.53 times more likely to have breast cancer?

A: No (not necessarily)

# Mammography

| Mammography result | Cancer status (biopsy) | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | TP = 29 | FP = 19 | 48 |
| Negative | FN = 1 | TN = 11 | 12 |
| Total | 30 | 30 | N = 60 |

LR+ = (29/30) / (19/30) = 1.53

Of those with positive tests 29/19 have cancer (1.53)

# Mammography

| Mammography result | Cancer status (biopsy) | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | TP = 29 | FP = 1881 | 1910 |
| Negative | FN = 1 | TN = 1089 | 1090 |
| Total | 30 | 2970 | N = 60 |

LR+ = (29/30) / (1881/2970) = 1.53

Of those with positive tests 29/1881 have cancer (0.02)

# Misinterpretation

People often make the mistake of thinking that if a test is 90% sensitive, then a positive test means that there is a 90% chance of having the disease.

This is known as the “Prosecutors fallacy” or the error of the transposed conditional

# Positive predictive value (PPV)

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | **TP** | FP | TP + FP |
| Negative | | | |
| Total | | | |

# Negative predictive value (NPV)

| Test Result | True condition/Disease status | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | | | |
| Negative | FN | **TN** | TN + FN |
| Total | | | |

# Mammography (population 1)

| Mammography result | Cancer status (biopsy) | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | TP = 29 | FP = 19 | 48 |
| Negative | FN = 1 | TN = 11 | 12 |
| Total | 30 | 30 | 60 |

PPV: 29/48 = 0.60

NPV: 11/12 = 0.92

# Mammography (population 2)

| Mammography result | Cancer status (biopsy) | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | TP = 29 | FP = 1881 | 1910 |
| Negative | FN = 1 | TN = 1089 | 1090 |
| Total | 30 | 2970 | N = 60 |

PPV: 29/1910 = 0.02

NPV: 1089/1090 = 0.99

# Beyond Basic Statistics for DA.

- Parametric models and smooth ROC curves
- Optimal thresholds
- Summary ROC curves (SROC) – meta analysis of DA studies
- Much more…….

The examples in this talk are taken from this book.