Learning Objectives

1. Learn how to calculate basic accuracy statistics such as sensitivity, specificity, likelihood ratios and AUC
2. Understand reasons for differences in diagnostic accuracy: real differences, bias, random variation, cut-offs.
3. Understand the difference between tests conducted under ideal conditions vs real conditions
4. Understand the role of higher-level approaches to performance evaluation
No Disclosures
Testing a Test: Beyond Sensitivity and Specificity:

Robert Schmidt MBA MD PhD MMed
Tests are Central to Medicine

• Diagnosis
• Prognosis
• Monitoring
• Management
Tests Exert Great Leverage

Medical Costs

3%

Lab Tests
Hierarchy of Effectiveness

- Societal Impact
- Cost effectiveness
- Clinical effectiveness
- Clinical performance
- Analytical performance
What this talk is about

Evaluating Tests:
• Accuracy
• Usefulness
• Test Comparisons
• Limitations
• Future Directions

Analytical performance
Clinical performance
Clinical effectiveness
Cost effectiveness
Societal Impact
Case:

Your father has just returned from his annual physical. His doctor suggested that he consider a prostatic specific antigen (PSA) test to screen for prostate cancer. He is unsure what to do and asks your advice. Should he take the test?
The basic task: classification

Test cutoff = 18

Test negative

No Disease

Test positive

Disease present

Test result
### Perfect Test:

<table>
<thead>
<tr>
<th>Test</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>200</td>
<td>0</td>
<td>200</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

### Disease vs. Test

- **No Disease**
- **Disease present**

![Graph showing distribution of test values with disease present and no disease]
Not so Perfect Test:

<table>
<thead>
<tr>
<th>Disease</th>
<th>Test</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>Positive</td>
<td>148</td>
<td>40</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>52</td>
<td>160</td>
<td>212</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Disease present

No Disease

Test value
<table>
<thead>
<tr>
<th>Test</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>148</td>
<td>40</td>
<td>188</td>
</tr>
<tr>
<td>Negative</td>
<td>52</td>
<td>160</td>
<td>212</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>Disease present</td>
<td>No Disease</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td><strong>Present</strong></td>
<td><strong>Absent</strong></td>
<td><strong>Total</strong></td>
</tr>
<tr>
<td>Positive</td>
<td>148</td>
<td>40</td>
<td>188</td>
</tr>
<tr>
<td>Negative</td>
<td>52</td>
<td>160</td>
<td>212</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>200</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>Disease</td>
<td>Present</td>
<td>Absent</td>
<td>Total</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Positive</td>
<td>148</td>
<td>40</td>
<td>188</td>
</tr>
<tr>
<td>Negative</td>
<td>52</td>
<td>160</td>
<td>212</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
<td>400</td>
</tr>
</tbody>
</table>
How well did we classify those with disease?

Sensitivity = 148/200 = 74%
How well did we classify those without disease?

Specificity = 160/200 = 80%
**Sensitivity** = accuracy in the **diseased** group

**Specificity** = accuracy in the **nondiseased** group
Two useful mnemonics*

SnNout:

High **S**ensitivity Test with a **N**egative result rules **out**

SpPin:

High **S**pecificity Test with a **p**ositive result rules **in**

*Guyatt & Sackett, Guide to EBM, 2002*
PSA test

- Sensitivity 90%
- Specificity 20%

How might this test be useful?
- SnNout?
- SpPin?
Results May Vary......

Sensitivity of PSA Tests
Specificity of PSA Test
Sensitivity and Specificity Depend on Cutoff Values

<table>
<thead>
<tr>
<th></th>
<th>Cutoff A</th>
<th>Cutoff B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>74%</td>
<td>22%</td>
</tr>
<tr>
<td>Specificity</td>
<td>80%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>Positive</td>
<td>148</td>
<td>40</td>
</tr>
<tr>
<td>Negative</td>
<td>52</td>
<td>160</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Sensitivity = 74%
Specificity = 80%
Threshold Effects on Test Performance

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>Positive</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>Negative</td>
<td>156</td>
<td>196</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Sensitivity = 22%
Specificity = 98%
Threshold Effects on Test Performance

<table>
<thead>
<tr>
<th>Disease</th>
<th>Test</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>196</td>
<td>156</td>
<td>312</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>4</td>
<td>44</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>200</td>
<td>200</td>
<td>400</td>
</tr>
</tbody>
</table>

Sensitivity = 98%
Specificity = 22%
Tradeoff: Specificity vs Sensitivity

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>98</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>74</td>
<td>80</td>
</tr>
<tr>
<td>20</td>
<td>22</td>
<td>98</td>
</tr>
</tbody>
</table>
## How to Compare Tests

<table>
<thead>
<tr>
<th>Specificity</th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>0.2</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>0.3</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>0.4</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>0.5</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>0.6</td>
<td>0.80</td>
<td>0.56</td>
</tr>
<tr>
<td>0.7</td>
<td>0.70</td>
<td>0.40</td>
</tr>
<tr>
<td>0.8</td>
<td>0.55</td>
<td>0.20</td>
</tr>
<tr>
<td>0.9</td>
<td>0.38</td>
<td>0.10</td>
</tr>
<tr>
<td>1.0</td>
<td>0.10</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Receiver Operating Characteristic (ROC) Curve

Sensitivity = TPR

1 – specificity = FPR

<table>
<thead>
<tr>
<th>Specificity</th>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>0.2</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>0.3</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>0.5</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>0.6</td>
<td>0.80</td>
<td>0.56</td>
</tr>
<tr>
<td>0.7</td>
<td>0.70</td>
<td>0.40</td>
</tr>
<tr>
<td>0.8</td>
<td>0.55</td>
<td>0.20</td>
</tr>
<tr>
<td>1.0</td>
<td>0.10</td>
<td>0.01</td>
</tr>
</tbody>
</table>
ROC Curve for The Perfect Test

Sensitivity = TPR

1 - specificity = FPR
Sensitivity = TPR

1 – specificity = FPR

Perfect Test

Useless Test*

*Heads = Positive
Tails = Negative
Test Performance is Related to Area Under the Curve (AUC)

Sensitivity = TPR

Perfect Test

Useless Test

Increasing Diagnostic Accuracy

$1 - \text{specificity} = \text{FPR}$
Area Under the Curve (AUC)

Perfect Test
AUC = 1.0

Real Test

Useless Test
AUC = 0.5
Does the AUC mean anything?
Does the AUC Mean Anything?

AUC = Average Sensitivity

Sensitivity = TPR

Perfect Test

Useless Test

AUC = Average Sensitivity

1 – specificity = FPR
Meaning of AUC

\[ \text{Prob} \left( T_{\text{Disease}} > T_{\text{No Disease}} \right) \]

\[ \text{AUC} = 1.0 \rightarrow \text{perfect separation} \]
Meaning of AUC

\[
\text{Prob} \left( T_{\text{Disease}} > T_{\text{No Disease}} \right)
\]

AUC = 0.50
No separation
# PSA vs PSA velocity

<table>
<thead>
<tr>
<th>Study</th>
<th>PSA velocity</th>
<th>PSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eggener, 2005</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>Ciatto, 2004</td>
<td>0.74</td>
<td>0.67</td>
</tr>
<tr>
<td>Berger, 2007</td>
<td>0.87</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Take home message:

Sensitivity
Specificity

Threshold Effects

ROC Curves
AUC

No Threshold Effects
How do I know if a test is useful?
Is this test useful?

Sensitivity = TPR

1 - specificity = FPR

Test A

AUC = 0.8
Usefulness is defined by the customer

Will this test tell me whether my patient has Prostate Cancer?
Usefulness is defined by the customer

The AUC is 0.75!

Huh?
The customer’s problem:

Uncertainty → TEST → Certainty
Usefulness = Capabilities - Requirements
How to think about certainty

The Odds Scale

\[
\text{Odds} = \frac{\text{Prob}(\text{Disease})}{\text{Prob}(\text{No Disease})} = \frac{P}{1-P}
\]

1:10 Low

1:1 Equal

10:1 High
Capability = Change in Certainty

\[
\frac{\text{Prob}(\text{Disease})}{\text{Prob}(\text{No Disease})}
\]

Test Impact

Odds of Disease
Before test

Change

Odds of disease
After test
What is the impact of a positive result?

Positive Likelihood ratio, LR+

\[ LR^+ = \frac{\text{Prob} (\text{Positive} | \text{disease})}{\text{Prob} (\text{Positive} | \text{no disease})} = \frac{40/200}{4/200} = 10.0 \]

Bigger LR+ is better.
What is the impact of a *negative* result?

Negative Likelihood ratio, LR-:

\[
LR^- = \frac{\text{Prob}(\text{Negative} \mid \text{disease})}{\text{Prob}(\text{Negative} \mid \text{no disease})} = \frac{160/200}{196/200} = 0.8
\]

Smaller LR- is Better
Key Relationship

Posterior Odds $= LR \times$ Prior Odds
Likelihood = Impact Factor

\[
\frac{\text{Prob(Disease)}}{\text{Prob(No Disease)}}
\]

Prior Odds \times LR = Posterior Odds
Two ways to be certain:

Prior Odds

Posterior Odds

Prob(Disease) / Prob(No Disease)

1:10

1:1

10:1

Exclude

Confirm

$\text{Prior odds} \times \text{Likelihood Ratio} = \text{Posterior odds}$
A test can solve the problem if:

\[ LR_{\text{Actual}} > LR_{\text{Required}} \]

\[ LR^{+} > \frac{PosteriorOdds}{PriorOdds} \quad \text{Confirm} \]

Or:

\[ LR^{-} < \frac{PosteriorOdds}{PriorOdds} \quad \text{Exclude} \]
Futile test: \textbf{Required} \ LR^+ = \frac{\text{PostOdds}}{\text{PriorOdds}} = \frac{10}{2} = 5

I have a slight hunch that this person may have Prostate CA. I will do a biopsy if the odds in favor of disease are around 10:1.
Rule-in, Confirmation Zone

High LR+ is better

Sensitivity = TPR

1 - specificity = FPR

Required $LR^+ > \frac{Posterior\ Odds}{Prior\ Odds} = 5$
Rule-out, Exclusion Zone:
Low LR- is better

Sensitivity = TPR

Required LR- < \( \frac{\text{Posterior Odds}}{\text{Prior Odds}} = 0.2 \)

1 – specificity = FPR
Capabilities > Requirements?

Sensitivity = TPR

1 – specificity = FPR

SnNout
SpPin

Required (LR), determined by the customer

actual LR, capabilities
Key Points:

• Accuracy ≠ Usefulness
• Potential Usefulness = LR = f(Sn,Sp)
• Usefulness = Capabilities - Requirements:
  – The objective (exclude, confirm)
  – Prior uncertainty
  – Required certainty
  – The Test Impact (actual LR vs required LR)
Setting Test Thresholds

Disease present

No Disease

??????
ROC curve = set of available Likelihood Ratios

Sensitivity = TPR

1 - specificity = FPR

Low LR+ = 1
No impact

High LR+ Confirm
How to select a threshold value

Sensitivity = TPR

$1 - \text{specificity} = \text{FPR}$
Diagnostic Zones for Thresholds

Sensitivity = TPR

1 − specificity = FPR

Deadly, easily curable disease
Pneumonia

Awful treatment, low chance of cure
Cancer
Accuracy, Usefulness and Optimality

Sensitivity = TPR

Requirements, determined by the customer

1 – specificity = FPR
Key Points – Setting Thresholds

• **Comparing Tests**
  – Thresholds are a nuisance
  – ROC/AUC facilitates comparisons of *diagnostic* accuracy

• **Using Tests**
  – Thresholds are required
    • Define a test
    • Link capabilities and requirements
    • Can be set to optimize performance
      – Optimum is context dependent
      – Depends on error costs
Comparing Test Performance

Why do test results differ?

1. True differences
2. False differences
   - (bias)
   - thresholds
3. Random variation (imprecision)
Evaluating Test Accuracy: Ideal vs Realistic Conditions

Consecutive Patients → Index Test → Gold Standard

Ideal

Realistic
Test Performance
Ideal vs Realistic Conditions

1 - specificity = FPR

Sensitivity = TPR
Referral Pattern & Disease Spectrum

Patient with Complaint

General Practitioner

Emergency Dept
Referral Pattern & Disease Spectrum

Patient with Complaint → General Practitioner → Specialist
Effect of Prior Testing
Will the index test perform differently?

*Index Test = the test of interest
## Defining a Test: PICCO

Sources of **Real** Differences: Context is Everything

| P | Population | Setting  
Exclusion/Inclusion criteria  
Referral pattern  
Comorbidities  
Age, Gender |
|---|------------|---------|
| I | Index Test | Method (in detail)  
Cutoff  
Skill level |
| C | Condition | Disease of interest |
| C | Comparator  
(reference test) | Definition of disease |
| O | Outcome measure | Diagnostic accuracy  
Discomfort, adverse events  
Operational (TAT, Availability, cost, etc) |
Comparing Test Performance

Why do test results differ?

1. True differences
2. False differences
   - bias
   - thresholds
3. Random variation (imprecision)
Sources of Bias
(Phantom Differences)

• Imperfect gold standard
• Verification Bias
• Indeterminate results
• Others....
Imperfect Gold Standard
(Differential verification)

Patients → Index Test → Negative X % → Brass Standard

Patients → Index Test → Positive X % → Gold Standard
Verification bias
(Differential sampling)

Patients → Index Test → Negative: X % → Gold Standard
Patients → Index Test → Positive: Y % → Gold Standard
Bias due to indeterminates

Evaluator A: No indeterminates
Low sensitivity
Low specificity

Evaluator B: Many indeterminates
High sensitivity
High specificity
### Indeterminates: Where do these values go?

<table>
<thead>
<tr>
<th>Index Test</th>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease Present</td>
</tr>
<tr>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Indeterminate</td>
<td>U</td>
</tr>
<tr>
<td>Negative</td>
<td></td>
</tr>
</tbody>
</table>
Comparing Test Performance

Why do test results differ?

1. True differences
2. False differences (bias)
3. Random variation (imprecision)
Understanding Statistical Variation in Studies
Meta-Analysis

Outlier?

Average

Specificity
## Comparing tests

<table>
<thead>
<tr>
<th>Source of Difference</th>
<th>Countermeasures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Differences</strong></td>
<td>Complete Reporting</td>
</tr>
<tr>
<td></td>
<td>PICCO</td>
</tr>
<tr>
<td></td>
<td>Meta-analysis</td>
</tr>
<tr>
<td><strong>False Differences</strong></td>
<td>Improved Study Design</td>
</tr>
<tr>
<td>bias</td>
<td>ROC Curves</td>
</tr>
<tr>
<td>thresholds</td>
<td></td>
</tr>
<tr>
<td><strong>Random variation</strong></td>
<td>Study design</td>
</tr>
<tr>
<td></td>
<td>Meta-analysis</td>
</tr>
</tbody>
</table>
Higher Levels of Test Evaluation

- Societal Impact
- Cost effectiveness
- Clinical effectiveness
- Clinical performance
- Analytical performance
Problems with Test Evaluation

• Potentially useful ≠ Clinically useful
• Potential problems
  – Tests are not used properly
  – Tests do not change diagnosis
  – Tests do not change management
• Tests are not used in isolation
  – Incremental value
Clinical Trial Evaluation of Tests

Key Question:

*Do patients who receive this test have better outcomes?*
Tests don’t exist in isolation

• Test Research vs Diagnostic Research
• What is the *incremental* value of a test?

Patient with Nausea → Test A → Test B → Test C → Index Test → Diagnosed
Tests are often combined

Test 1 \rightarrow \text{Probability of Disease} \\
Test 2 \\
Test n \\

Probability of Disease
The acid test
## Clinical Trial Evaluation

### Prostate Screening (PSA)

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Event Rate per 1000</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Screen</td>
<td>Screen</td>
</tr>
<tr>
<td>All cause mortality</td>
<td>200</td>
<td>198</td>
</tr>
<tr>
<td>Death from prostate CA</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Prostate CA diagnosis</td>
<td>44</td>
<td>64</td>
</tr>
</tbody>
</table>
Levels of Evaluation

**Therapeutics**

**Phase II/III Trial – Explanatory Trial**
- Scientific Perspective
  - Hypothesis: Does this drug affect outcomes?
- As-Treated Analysis
  - Carefully controlled population, setting
  - Carefully controlled administration and monitoring

**Phase III Trial – Pragmatic Trial**
- Policy Perspective
  - Hypothesis: Does prescribing this drug affect outcomes?
- Intention-to-Treat Analysis
  - Patients seeking treatment for condition
  - Usual conditions

**Diagnostics**

**Scientific Test Evaluation**
- Single test
  - Idealized population
  - Expert administration
  - Expert interpretation

**Pragmatic Test Evaluation**
- Multiple tests
  - Actual population
  - Usual conditions

---

**Analytical Performance**
- Discrimination
  - Ideal (disease vs non-disease)
- Cost Effectiveness
Cost-Effectiveness Plane

New Cancer Drug
CER = 100,000/QLY

New Diagnostic Test
CER = $10,000/QLY
Summary

• Many ways to assess performance
• Many reasons why studies differ
  – Real differences (PICCO)
  – False differences
    • Thresholds
    • Bias
  – Statistical variation
• Progress in Performance Evaluation
  – Quality of Reporting
  – Quality of studies
  – Types of studies
• Educating Clinicians
Testing A Test:
Beyond Sensitivity and Specificity