Study Design and Analysis: Diagnostic Studies

Mary Dunbar MD MSc MSc FRCPC
Assistant Professor of Pediatrics, University of Calgary
Pediatric Neurologist, Alberta Children’s Hospital

Slide Credit: many slides taken from Dr. S. Greenaway’s lecture from last year
Objectives

• Understand the issues in the evaluation of a diagnostic test
• Appreciate components of evaluating test performance
• precision and accuracy
• sensitivity and specificity
  • likelihood ratio
  • positive and negative predictive values
  • receiver operating characteristic (ROC) curves
• additional factors: cost, availability, acceptability, utility
Examples of Diagnostic Tests

- **Biochemical**
  - electrolytes, urea, creatinine

- **Imaging**
  - CXR, MRI

- **Genetic**
  - karyotype, array, WES

- **Microbiological**
  - blood culture

- **Physiological**
  - PFTs, exercise test, GTT

- **Clinical**
  - Lever sign to diagnose ACL tear

- **Patient-reported outcome measures**
  - questionnaire of symptoms to diagnose IBD
Purpose of diagnostic tests

- Diagnose a disease or condition
  - TSH
  - echocardiogram
- Exclude a disease or condition
  - HbA1C
  - Troponin
- Estimate prognosis
  - LDL cholesterol
  - BRCA1 mutation
- Inform treatment decisions
  - PSA
  - karyotype
Factors Affecting Diagnostic Test Performance

• Prevalence of the disease in the population
• Spectrum of the disease
• Often dependent on other factors
  • part of diagnostic pathway
  • test results may not be independent
  • often depend on prior knowledge
• Gold standard
  • established test which confirms the diagnosis
Types of Studies to Evaluate a Diagnostic Test

• Precision (reproducibility)
  • intra-observer (amount of variation for a single observer)
  • inter-observer (variation between 2 or more observers)

• Accuracy
  • cohort
  • case-control

• Costs, Risks and Acceptability
  • prospective
  • retrospective

• Improvement of clinical outcome
  • RCT
  • case-control
Precision

- Reproducibility or repeatability
  - Agreement between repeated measures
- Intra-observer variability
  - Agreement with your previous interpretation
- Inter-observer variability
  - Agreement between observers
Accuracy

• Closeness of measurements to a specific value
  • To what extent does the test give the right answer
  • Requires a gold standard (definitive assessment)
• Measures of accuracy
  • Sensitivity and specificity
  • Positive and negative predictive values
  • Receiver operating characteristic (ROC) curve
  • Likelihood ratio
Sensitivity & Specificity

• Sensitivity
  • proportion of positive tests out of total disease
  • Given you have the disease, proportion that have a positive test ($T^+|D^+$)
  • correctly identified positives
  • true-positive rate
  • The probability that a person with the disease is classified correctly by the test

• Specificity
  • proportion of negative tests out of total non-diseased
  • Given you don’t have the disease, proportion that have a negative test ($T^-|D^-$)
  • correctly identified negatives
  • true-negative rate
  • The probability that a person without the disease is classified correctly by the test
Dichotomous Outcome and Test Result 2x2 Contingency Table

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td>Negative test</td>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>
Calculating Sensitivity and Specificity

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td>Negative test</td>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Stroke</th>
<th>No stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT = stroke</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>CT = no stroke</td>
<td>161</td>
<td>136</td>
</tr>
</tbody>
</table>

Sensitivity: true positives/all stroke = 56/217 = 26%
Specificity: true negatives/all without stroke = 136/139 = 98%

Sensitivity & Specificity: classification

• Sensitivity and Specificity tell you about misclassification errors
• Studies that display results as sensitivity and specificity are Validation Studies
  • Step 1: obtain a sample of people with and without a disease
  • Step 2: administer a test or procedure to classify them
  • Step 3: compare the results of the classification to a “gold standard” and construct a 4x4 table
Sensitivity and Specificity - Challenges

• Never consider these two parameters separately
  • Trade off between sensitivity and specificity
  • As one increases, the other decreases
  • e.g. higher cutoff leads to increased specificity but decreased sensitivity

• A highly sensitive test is prone to false-positives
  • incorrectly label someone as having the disease

• A highly specific test is prone to false-negatives
  • fail to identify disease

• What is important to you?
  • Avoid missing someone or avoid incorrectly labelling someone?
Trade off Between Sensitivity and Specificity
Trade off Between Sensitivity and Specificity

Aii

Bii

Aiii

Biii
Sensitivity and Specificity - Challenges

• Affected by severity of disease
  • results from a CXR for detection of lung cancer will depend on severity of illness and stage of the disease, size of the tumour etc.

• Sensitivity and specificity describe how well a test performs
  • Don’t convey significance of the test result for an individual patient
Likelihood Ratio

• Assesses potential utility of a diagnostic test
  • Assesses how likely the patient with a positive test has the disease

• Probability of positive test given disease relative to probability of positive test given no disease (true positives/false positives)

• Answers question: How much more likely is a positive test result in the presence of disease compared with absence of disease?

• LR = sensitivity/(1-specificity)

• Answer is an odds
Likelihood Ratio

• Has predictive value and stable with changes in prevalence
• Ranges from zero to infinity
• The higher the value, the more likely the patient has the condition
  • 0 - 1 = decreased evidence for disease
  • 1 = no diagnostic value
  • >1 = increased evidence for disease
## Likelihood Ratio Example

<table>
<thead>
<tr>
<th>Serum Ferritin (mg/dL)</th>
<th>LR (of iron deficiency anemia)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;15</td>
<td>51.8</td>
</tr>
<tr>
<td>15-24</td>
<td>8.8</td>
</tr>
<tr>
<td>25-34</td>
<td>2.5</td>
</tr>
<tr>
<td>45-100</td>
<td>0.5</td>
</tr>
<tr>
<td>&gt;100</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Sloane 2008
Prediction

• Predictive values

• Ability of a diagnostic test to make a diagnosis in the future

• Positive predictive value (PPV)
  • proportion of diseased with positive test result
  • proportion of people with a positive test who have the disease

• Negative predictive value (NPV)
  • proportion of healthy individuals with a negative test result
  • proportion of people with a negative test who are free of disease
Prediction

- A test with a high positive predictive value makes the disease quite likely in a subject with a positive test.
- A test with a high negative predictive value makes the disease quite unlikely in a subject with a negative test.

Positive predictive value (PPV) = true positive tests/all positive tests
Negative predictive value (NPV) = true negative tests/all negative tests
## Prediction

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td>Negative test</td>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Stroke</th>
<th>No stroke</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CT = stroke</td>
<td>56</td>
<td>3</td>
<td>59</td>
</tr>
<tr>
<td>CT = no stroke</td>
<td>161</td>
<td>136</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>217</td>
<td>139</td>
<td>356</td>
</tr>
</tbody>
</table>

Positive predictive value (PPV) = true positive tests/all positive tests
Negative predictive value (NPV) = true negative tests/all negative tests

Positive predictive value (PPV) = \(\frac{56}{59} = 95\%\)
Negative predictive value (NPV) = \(\frac{161}{297} = 54\%\)

Calculating Sensitivity and Specificity

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td>Negative test</td>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Stroke</th>
<th>No stroke</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CT = stroke</td>
<td>56</td>
<td>3</td>
<td>59</td>
</tr>
<tr>
<td>CT = no stroke</td>
<td>161</td>
<td>136</td>
<td>297</td>
</tr>
<tr>
<td></td>
<td>217</td>
<td>139</td>
<td>356</td>
</tr>
</tbody>
</table>

Sensitivity: true positives/all stroke = 56/217 = 26%
Specificity: true negatives/all without stroke = 136/139 = 98%

Predictive values - Challenges

• Cannot be used in case-control studies
  • used for random samples or cohorts where observed prevalence is equivalent to true prevalence

• Affected by prevalence (proportion of subjects with disease)
  • high prevalence
    • PPV increases and NPV decreases
  • low prevalence
    • PPV decreases, NPV increases

• Less portable from population to population
  • due to effect of prevalence
### Effect of Prevalence on PPV and NPV

<table>
<thead>
<tr>
<th>Prevalence</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.987097</td>
<td>0.400000</td>
</tr>
<tr>
<td>75%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.962264</td>
<td>0.666667</td>
</tr>
<tr>
<td>50%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.894737</td>
<td>0.857143</td>
</tr>
<tr>
<td>25%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.739130</td>
<td>0.947368</td>
</tr>
<tr>
<td>10%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.485714</td>
<td>0.981818</td>
</tr>
<tr>
<td>1%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.079070</td>
<td>0.998319</td>
</tr>
<tr>
<td>0.01%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.0000849</td>
<td>0.999983</td>
</tr>
<tr>
<td>0.001%</td>
<td>0.85</td>
<td>0.9</td>
<td>0.0000085</td>
<td>0.999998</td>
</tr>
</tbody>
</table>

For a test with 85% sensitivity and 90% specificity:

- **PPV** decreases as the prevalence decreases.
- **NPV** increases as the prevalence decreases.
Effect of prevalence

• Example of newborn screening for congenital hypothyroidism
• Amazing test
• But low prevalence = low PPV

<table>
<thead>
<tr>
<th></th>
<th>Cord sampling</th>
<th>Heel-stick sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Specificity</td>
<td>99.6%</td>
<td>98.3%</td>
</tr>
<tr>
<td>Recall rate</td>
<td>0.04%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>7.95%</td>
<td>2.30%</td>
</tr>
</tbody>
</table>

https://www.researchgate.net/publication/336248581_Cord_blood_versus_heel-stick_sampling_for_measuring_thyroid_stimulating_hormone_for_newborn_screening_of_congenital_hypothyroidism/figures?lo=1
Prevalence and Diagnostic Tests

• Diagnostic tests function best when prevalence is between 40-60%
  • Chose the right population to test

• **Function poorly at extremes of prevalence**

• “When you are already pretty sure that the patient either does or does not have the diagnosis in question, additional testing may not alter that probability very much”

• e.g. ECHO for endocarditis or chest CT for pulmonary embolus
Summary of terms

• Sensitivity and specificity
  • How good is the test compared to gold standard?

• Likelihood ratio
  • How much more likely is a positive test result in the presence of disease compared with absence of disease? (true positives/false positives)

• Predictive value
  • Given a test result, what is the probability of actually having the disease?
Receiver Operating Characteristic (ROC) Curves

- Test result is not simply positive or negative
- Continuous test results
- Potentially multiple cutoffs
- Sensitivity (Y-axis) vs. 1-specificity (X-axis)
- Best cut-off maximizes sensitivity and specificity
  - 1 = perfect test
  - 0.5 = useless test (equivalent to random chance)
- Quantifies information gain for a test
- Provides summary estimate of the accuracy of the test
Area Under the ROC Curve (AUC)

- Values between 0.0 and 1.0
  - perfectly inaccurate to perfectly accurate
  - 0.5 = useless test

<table>
<thead>
<tr>
<th>AUC values</th>
<th>Test quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9–1.0</td>
<td>Excellent</td>
</tr>
<tr>
<td>0.8–0.9</td>
<td>Very good</td>
</tr>
<tr>
<td>0.7–0.8</td>
<td>Good</td>
</tr>
<tr>
<td>0.6–0.7</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>0.5–0.6</td>
<td>Unsatisfactory</td>
</tr>
</tbody>
</table>
Examples of ROC Curves

Ai All study subjects

Bi Controls vs. mild CP

Ci Controls vs. severe CP

Aii 0.74

Bii 0.72

Cii 0.81
Additional Considerations

• Cost
• Availability
• Acceptability
  • i.e. invasive test with potentially serious complications
• Clinical utility
  • ideally assessed using a RCT
    • assess outcomes
    • document adverse events
    • assess impact on decision-making
    • assess patient satisfaction and cost-effectiveness
Example

- Evaluation of a potential screening test for Cerebral Palsy
Why a new test?

• Cerebral palsy is an impairment of motor development due to a static abnormality of the CNS that occurs before the age of 1 (ie, in development)
• Affects ~1/500 children
• CP is a clinical diagnosis
• CP takes time to become apparent due to maturation of the CNS
• Early interventions improve outcomes
• How can we identify children at risk?
  • Term infants with encephalopathy at birth ~12% develop CP
Classic risk factors

• Prematurity (~40%)
• Bad delivery (~10-20%)
• These children are easy to identify and follow
• But these account for a minority of CP cases (~50%)
• What about the rest?
Study

• Canadian Cerebral Palsy registry = cases = 1265
• APrON (Alberta Pregnancy Outcomes and Nutrition) = controls = 1985
• Look a common elements and try to find ones specific to CP
Some figures removed as do not have permission to share them (article in press)

They will be available in ~December or so, feel free to contact me if you’d like them mary.Dunbar@ahs.ca
CAVEAT

- The "prevalence" of CP in our study is high! 38%
- This means PPV and NPV are very misleading if we look at the general population! (~0.2%)
  - Recall the PPV and NPV should not be used in a case control study
  - (doesn’t stop the reviewers from asking for it)
Is this acceptable??

• Screening test – want high sensitivity, low specificity
• But low specificity = worried parents, unnecessary tests
• **Acceptability**: Screening is non-invasive (no blood, etc)
• **Availability**: can be done by anyone, most variables will be known
• **Utility**: does this actually identify additional cases of CP???
• **Cost**: tool is free, but requires time; next level screening requires resources
• Next level screening non-invasive (well baby check)
• Tiny subset referred for more intensive screening such as Hammersmith Infant Neurological Examination, General Movements Assessment (can be administered by PTs)
Summary

• Multiple metrics to evaluate a diagnostic test
• Test performance
  • precision (reproducibility) and accuracy
  • sensitivity or specificity
  • likelihood ratio
• Positive and negative predictive values
  • affected by disease prevalence
  • function poorly at the extremes
• ROC curves
  • estimate accuracy of the test for different cutoff values
  • summarized with AUC
• Impact and non-clinical factors
Thanks!

Mary.Dunbar@ahs.ca