How to Design a Diagnostic Accuracy Study

Chad Cook PhD, PT, MBA, OCS, FAAOMPT
Professor and Chair
Walsh University

Profound Statement

• Exaggerated reports of diagnostic utility from poorly designed studies can potentially result in an inaccurate diagnosis, inappropriate treatment, premature adoption of a special test that provides little value, and errors in clinical decision-making.


Our Findings

• 3.1% of tests obtained a utility score of 1 in our textbook for Utility.
• Utility = 1 means “Evidence Strongly Supports the Use of this Test”

Bad Research

• Although I think that, in many cases, the research on special tests is more to blame than the tests themselves and that these tests should remain part of a skilled clinical examination, I agree with the message: JMNT and other journals need high-quality diagnostic studies since the components of a physical examination influence clinical decision making.


In truth, Most Published Diagnostic accuracy studies.....

Because it isn’t easy

\[
\begin{align*}
e^{j\theta} &= f = 1 \angle 0^\circ \\
e^{-j\theta} &= f = 1 \angle -90^\circ \\
e^{2j\theta} &= f \pm 180^\circ = -1 \\
f &= f' = j \\
\frac{1}{f} &= j \\
\frac{1}{f} &= -j \\
-15 &= 15 \angle -90^\circ
\end{align*}
\]
Ingredients for a Diagnostic Accuracy Study

- Acceptable Reference Standard
- Case Based, Case Control
- Dedicated tests and measures (and an acceptable operational definition)
- An appropriate spectrum
- An appropriate sample (and size)
- Blinded researchers
- Calculation of Metrics
- STARD checklist

Acceptable Reference Standard

- Gold Standard? (Unquestionable method to confirm a diagnosis)
- Reference Standard? (“Best” selection of the current methods for confirming a diagnosis)
- Criterion Standard? (The best proxy method to represent a diagnosis)

Reference Standard

- Surgical Confirmation (top shelf)
- Imaging Confirmation of a clinical finding
- Imaging Finding
- Clinical Finding (change in condition)
- Expert Opinion

Current Reference Standards?

- Cervical radiculopathy
- Herniated lumbar spine
- Knee meniscus tear
- Rotator Cuff Tear
- Osteoarthritis
- Lumbar spine stenosis
- Cervical myelopathy

What you will Need

- Imaging
- Surgical Confirmation

Case Control Bias?

Dedicated tests and measures (and an acceptable operational definition)

• How do you do McMurray’s Test?

Validity of the McMurray’s Test and Modified Versions of the Test: A Systematic Literature Review

How do you do a Straight Leg Raise?

How do you do the Active Straight Leg Raise


An Appropriate Spectrum

• A “Spectrum” is the characteristic and representation of a population
• Thus, the sample should reflect that population


Who is at Fault?

Problems?
An appropriate sample (and size)

- Sample should be calculated based on sensitivity findings (if screening) and specificity (if confirmation)
- Should be adjusted for variance
- Should be calculated for conditions


How do we Do?

- Of 43 studies on diagnostic accuracy, the median sample size was 118 (interquartile range 71-350) and the median prevalence of the target condition was 43% (27-61%).
- Median number of patients with the target condition—needed to calculate a test’s sensitivity—was 49 (28-91).
- Two of the 43 studies (5%) reported a priori calculations of sample size.
- Twenty articles (47%) reported results for patient subgroups. No studies reported that sample size was calculated on the basis of preplanned analyses of subgroups.

Bachmann et al. Sample sizes of studies on diagnostic accuracy: literature survey BMJ 2006; 332 : 1127

Blinded Researchers

- Those that use the index test need to be blinded to the reference standard
- Those that define the reference standard need to be blinded from the index tests
- The index tests cannot be part of the reference standard (incorporation bias)

Problems with This?

- Bias
- Self promoting studies
- Bogus metrics
- Self fulfilling prophecies

Calculation of Metrics

- Reliability
- Sensitivity
- Specificity
- Likelihood ratios
- Positive and Negative Predictive Values
- Accuracy

Purpose of a Diagnostic Accuracy Study

- Determine the “accuracy” or a specific test or a cluster of tests and measures toward the 1) diagnosis, 2) prognosis, or 3) best intervention method
- Accuracy is scored from 0 to 100% and is calculated \( \frac{(TP+TN)}{(TP+TN+FN+FP)} \)
Accuracy = (TP+TN)/(TP+TN+FN+FP)

Example

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TP = 56</td>
<td>FP = 12</td>
</tr>
<tr>
<td>FN = 22</td>
<td>TN = 51</td>
</tr>
</tbody>
</table>

66 + 51 / 56 + 51 + 22 + 12 = 107 / 141 or 76%

Contradictory Statement?

"Tests can have low accuracy but may still provide value during the diagnostic process"

Reliability

- Reliability or chance corrected agreement is the estimate to which a test score is free from error and if erroneous the degree in which the value varies from a true score.
- 0.0 to 0.4 = Poor
- 0.4 to 0.6 = Fair
- 0.6 to 0.8 = Good
- 0.8 to 1.0 = Exceptional
  
  Measured as a Kappa (dichotomous) or measured as an ICC (continuous)

Language Standards

- **Sensitivity**: Percentage of people who test positive for a specific disease among a group of people who have the disease
- **Specificity**: Percentage of people who test negative for a specific disease among a group of people who do not have the disease

Sensitivity Example

- 50 patients with arm pain associated with cervical radiculopathy
- Test was positive in 40 of the 50 cases
- Sensitivity = 40/50 or 80%
- Correct 80% of the time in cases that were cervical radiculopathy

http://www.ringgerpointbook.com/tfrasp2.gif
Specificity Example
- 50 patients with no arm pain associated with a cervical strain
- Test was positive in 5 of the 50 cases
- Specificity = 45/50 or 90%
- Correct 90% of the time in cases that were NOT cervical radiculopathy

Positive and Negative Predictive Value
- Positive predictive value: the proportion of patients with positive test results who are correctly diagnosed
- Negative predictive value: the proportion of patients with negative test results who are correctly diagnosed

Likelihood Ratios
- A high LR+ influences post-test probability with a positive finding
- A value of ≥1 rules in a diagnosis
- A low LR- influences post-test probability with a negative finding
- A value closer to 0 is best and rules out

For the Math Geeks
- Positive Likelihood Ratio = (sensitivity)/(1-specificity)
- Negative Likelihood Ratio = (1-sensitivity)/(specificity)

Influencing Decision Making

<table>
<thead>
<tr>
<th>LR</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;10</td>
<td>Large and often conclusive increase in the likelihood of disease</td>
</tr>
<tr>
<td>5-10</td>
<td>Modest increase in the likelihood of disease</td>
</tr>
<tr>
<td>2-5</td>
<td>Small increase in the likelihood of disease</td>
</tr>
<tr>
<td>1-2</td>
<td>Minimal increase in the likelihood of disease</td>
</tr>
<tr>
<td>1</td>
<td>No change in the likelihood of disease</td>
</tr>
<tr>
<td>&lt;1</td>
<td>Large and often conclusive decrease in the likelihood of disease</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LR</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5-1.0</td>
<td>Minimal decrease in the likelihood of disease</td>
</tr>
<tr>
<td>0.2-0.5</td>
<td>Small decrease in the likelihood of disease</td>
</tr>
<tr>
<td>0.1-0.2</td>
<td>Moderate decrease in the likelihood of disease</td>
</tr>
<tr>
<td>&lt;0.1</td>
<td>Large and often conclusive decrease in the likelihood of disease</td>
</tr>
</tbody>
</table>

- LR+ = Positive Likelihood ratio
- LR- = Negative Likelihood ratio

STARD Checklist
- In 1999, the Cochrane Group met and created the Standards for Reporting of Diagnostic Accuracy (STARD) checklist, a 25-item checklist
- The STARD checklist is designed to provide researchers with a checklist and flow diagram that should be followed for optimal study design.
STARD checklist

• Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.
• Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the (evaluated) index tests or the (golden) reference standard?

• Describe participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.
• Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?

• Describe the reference standard and its rationale.
• Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.
• Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.

• Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.
• Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.
• Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).
• Report when study was done, including beginning and ending dates of recruitment.
• Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centers).

• Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).

• Report time interval from the index tests to the reference standard, and any treatment administered between.

• Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.

• Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.

• Report any adverse events from performing the index tests or the reference standard.

• Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).

• Report how indeterminate results, missing responses and outliers of the index tests were handled.

• Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.

• Discuss the clinical applicability of the study findings.

• Retrospective Assessment
  • (Quality Assessment of Diagnostic Accuracy Studies)
  • 14 items (scored 0 to 14)
    – 1. Appropriate selection of patient spectrum
    – 2. Appropriate reference standard
    – 3. Absence of review bias (both test and diagnostic)
    – 4. Clinical review bias
    – 5. Reporting of ininterpretable/ indeterminate/ intermediate results

Whiting et al. BMC Medical Research Methodology. 2003, 3:25
• 10/14 or higher affects scoring


Bias

• Incorporation Bias
• Expectancy Bias
• Verification Bias
• Reference Bias
• Index performance Bias

Even when you design a good study.....

Questions?

Thank You!