Diagnosis: Interpreting Evidence in the absence of a reference standard - two case studies from national guidelines

Alfred Sackeyfio Presenter
Faisal Siddiqui
Abitha Senthinathan
Beth Shaw

GIN 2010

Shorts Clinical Guidelines Centre for Clinical Practice
Diagnosis of Latent tuberculosis

Tuberculosis: interferon gamma tests for the diagnosis of latent tuberculosis (Update)
DIAGNOSTIC TESTS

- Usually comparing the utility or accuracy of one or more index tests with that of a Reference standard.
- Typically the tests would assess measures such as sensitivity, specificity, positive and negative predictive values.
- Diagnostic technologies may be used for various purposes, including initial diagnosis, clinical monitoring, screening, treatment triage, assessing stages of disease progression and risk stratification.
Classic two by two table

- Sensitivity: Defined as the number of participants who correctly tested positive for the condition as a proportion of individuals who actually have the condition.
- This is given by \( a/(a+c) \)

<table>
<thead>
<tr>
<th>INDEX TEST</th>
<th>CONTESTED</th>
<th>NO CONDITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITIVE</td>
<td>TRUE POSITIVES (a)</td>
<td>FALSE POSITIVES (b)</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>FALSE NEGATIVES (c)</td>
<td>TRUE NEGATIVES (d)</td>
</tr>
</tbody>
</table>
**Classic two by two table**

- Specificity: Defined as the number of participants who correctly tested negative for the condition as a proportion of individuals who actually do not have the condition.
- This is given by \( \frac{d}{(b+d)} \)

<table>
<thead>
<tr>
<th>INDEX TEST</th>
<th>REFERENCE TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CONDITION</td>
</tr>
<tr>
<td>POSITIVE</td>
<td>TRUE POSITIVES (a)</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>FALSE NEGATIVES (c)</td>
</tr>
</tbody>
</table>
LATENT TB INFECTION

- LATENT DISEASE
  - NO SYMPTOMS
  - NOT TRANSMISSIBLE
  - SO WHY WORRY??

 Submission Number 2462
MAIN ISSUES

• Current practice?? TST
• Flaws with TST
  – Specificity results confounded by BCG vaccination and environmental non tuberculosis causing mycobacterium so potential for false positives
  – In immunocompromised patients, sensitivity results confounded by inability of immune system to develop antibodies to TST
  – Hence TST is not a good reference standard.
Index test

- Interferon Gamma Release Assay (IGRA)/(IGT)
- Numerous Studies published suggesting IGRA may be superior to TST.
- Published papers acknowledged flaws in design and quality of studies
- IGRA did not suffer from BCG and environmental mycobacteria issues that affects TST
- ........But IGRA may not be superior to TST in detecting latent infection in immunocompromised.
Index test (IGRA)

• Studies demonstrated the hypothesized superiority of IGRA over TST by showing that IGRA correlated more closely with the level of exposure to MTB than TST.
• This approach cannot directly measure the sensitivity and specificity of either assay for LTBI, but it can enable us to rank the tests according to their diagnostic utility.
• False positive rates and false negative rates cannot be determined.
• Hence in certain populations decisions on which test to recommend has become difficult. (E.g. children and compromised population)
RESULTS SHOWING SUMMARY STATISTICS COMPARING TST with IGT IN CHILDREN

Forest plot of meta-analysis of IGT and TST results based on high-risk and low-risk exposure

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>log(Odds Ratio)</th>
<th>SE</th>
<th>Total</th>
<th>Control</th>
<th>Total</th>
<th>Weight</th>
<th>Odds Ratio IV, Random, 95% CI</th>
<th>Odds Ratio IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brock et al 2004(1434)</td>
<td>0.6576</td>
<td>0.7124</td>
<td>85</td>
<td>85</td>
<td>13.3%</td>
<td>1.93 [0.48, 7.80]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chun et al 2003 (276)</td>
<td>0.1137</td>
<td>0.6096</td>
<td>71</td>
<td>71</td>
<td>16.3%</td>
<td>1.12 [0.34, 3.70]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hansted et al 2009 (3427)</td>
<td>0.9399</td>
<td>0.5417</td>
<td>97</td>
<td>97</td>
<td>18.9%</td>
<td>2.56 [0.89, 7.40]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higuchi et al 2009(164)</td>
<td>1.4986</td>
<td>0.6008</td>
<td>313</td>
<td>306</td>
<td>16.7%</td>
<td>4.48 [1.38, 14.53]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lighter et al 2009 (282)</td>
<td>2.3114</td>
<td>0.6039</td>
<td>174</td>
<td>174</td>
<td>16.5%</td>
<td>10.09 [3.09, 32.95]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Okada et al 2008(393)</td>
<td>0.7371</td>
<td>0.556</td>
<td>195</td>
<td>195</td>
<td>18.3%</td>
<td>2.09 [0.70, 6.21]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total (95% CI) 935 928 100.0% 2.86 [1.56, 5.23]

Heterogeneity: Tau² = 0.21; Chi² = 7.94, df = 5 (P = 0.16); I² = 37%
Test for overall effect: Z = 3.40 (P = 0.0007)
GDG Decision making

- Discussions revolved around the inability to determine false positive and false negative rates.
- The populations which posed the most challenge were children and the immunocompromised populations.
- GDG felt the implications of a misdiagnosis in these populations would be massive.
  - To diagnose a child with LTBI would mean being put on medications which would induce hepatototoxicity.
  - To miss a diagnosis in a child with LTBI implies the child could potentially go on and develop tb in the brain.
Motor neurone disease: the use of non-invasive ventilation in the management of motor neurone disease

- What is the diagnostic accuracy of specific investigations to confirm, assess and monitor the severity of respiratory impairment in patients with MND?
Population

- Adults (aged 18 and over) with a diagnosis of MND in primary care and community settings, secondary care and tertiary care.

**Subgroups:**
- People with MND who have moderate or severe bulbar impairment
- People with MND who have severe cognitive impairment or dementia

- Any diagnostic/investigation tools used to confirm, assess and monitor respiratory impairment in people with motor neurone disease.
  - Diagnostic studies with appropriate reference standard and statistical analysis that demonstrated diagnostic accuracy.
Specific considerations

- The five included studies were appraised and presented using GRADE profiles adapted for diagnostic tests or strategies (Schunemann et al. 2008).
- GDG members agreed the following cut-offs for sensitivity, specificity, positive and negative predictive value:
  - > 90% = very good
  - 70–90% = good
  - 60–69% = reasonably good
  - < 60% = poor.

**BUT**

- For review question (diagnostic accuracy), there is no single reference standard (gold standard) for respiratory impairment/insufficiency, hence, unable to draw conclusion from the evidence which is the best investigation tool.
- The five included studies used different reference standards to define respiratory impairment.
- There was an absence of clear evidence about which respiratory function tests are best for detecting respiratory impairment.
GDG Decision making

- The GDG acknowledged that it was particularly challenging to draw conclusions about which respiratory function tests to recommend based on the available evidence, because different reference standards were used to define respiratory impairment in the included studies.
- GDG therefore came to the consensus that only high-quality evidence should be used as the basis for developing recommendations.
- The GDG also agreed that, because of the potentially fatal consequences of not detecting respiratory impairment (that is, sudden respiratory failure, complications or death), respiratory function tests with high sensitivity (rather than high specificity) are most important.
- The GDG recommended standard tests for detecting respiratory impairment in current UK practice and using the clinical utility of the tests as well.
- Hence, tests with high quality evidence but not commonly used in the UK and have poor clinical utility were not recommended.
- The cut off values were determined using the GDG’s clinical experience amongst them.
FINDINGS

- Reliance on surrogate measures of effect are helpful to estimate the utility of the test.
- In certain situations there are a few reference standards and there is more than one standard measure of effect.
- However, the vital detailed information is not determined and in cases where it is determined there may be several values of one measure to choose from.
- This makes it difficult for users of the diagnostic tools to rely on them for a decision.
- The useful but limited information that surrogate measures give should be complemented with clinical expertise and epidemiology of the condition.
CONCLUSIONS

- Secondary researchers should continue to explore innovative techniques to identify appropriate surrogate measures.
- Thorough knowledge of disease and its natural history with transition probabilities is required.
- Clinicians/Users of diagnostic tools should engage with secondary researchers to identify areas of research which would enhance a focused clinical decision in a very ambiguous area of health care.