Diagnostic research designs: an introductory overview

Madhukar Pai, MD, PhD
Associate Professor of Epidemiology, McGill University
Montreal, Canada
Email: madhukar.pai@mcgill.ca

Classical EBM approach to diagnosis: compute sens/spec, LRs, and work out the post-test probabilities...

Above this point, treat

Disease ruled IN

Below this point, no further testing

Disease not ruled in or out

Disease ruled OUT
Bayes' theory

- Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities.
- Every test is done with a certain probability of disease - degree of suspicion [pre-test or prior probability].
- The probability of disease after the test result is the post-test or posterior probability.

What you thought before + New information = What you think now

\[
\text{post-test probability} = \frac{\text{pre-test probability} \times \text{likelihood ratio}}{1 + \text{pre-test probability} / \text{likelihood ratio}}
\]

Post-test odds = Pre-test odds x Likelihood ratio

Bayesian approach to diagnosis

- An accurate test will help reduce uncertainty.
- The pre-test probability is revised using test result to get the post-test probability.
- Tests that produce the biggest changes from pretest to post-test probabilities are most useful in clinical practice [very large or very small likelihood ratios].
- LR also called "Bayes Factor"
The diagnostic process is Bayesian, probabilistic, multivariable and sequential

1. A diagnosis starts with a patient presenting a complaint (symptom and/or sign) suggestive of a certain disease to be diagnosed.
2. The subsequent work-up is a multivariable process. It involves multiple diagnostic determinants (tests) that are applied in a logical order: from age, gender, medical history, and signs and symptoms, to more complicated, invasive, and costly tests.
3. Setting or ruling out a diagnosis is a probabilistic action in which the probability of the presence or absence of the disease is central. This probability is continuously updated based on subsequent diagnostic test results.
4. The true diagnostic value of a test is determined by the extent to which it provides diagnostic information beyond earlier tests, that is, materially changes the probability estimation of disease presence based on previous test results.
5. The goal of the diagnostic process is to eventually rule in or out the disease with enough confidence to take clinical decisions. This requires precise estimates of the probability of the presence of the target disease(s).


Some differences

- Test research vs. diagnostic research
- Diagnosis vs. screening
- Diagnosis vs. prediction
Diagnosis Vs Screening

- A diagnostic test is done on sick people
  - patient presents with symptoms
  - pre-test probability of disease is high (i.e. disease prevalence is high)
- A screening test is usually done on asymptomatic, apparently healthy people
  - healthy people are encouraged to get screened
  - pre-test probability of disease is low (i.e. disease prevalence is low)
Diagnosis vs. prediction

- **Diagnosis:**
  - Disease has already occurred and we are trying to detect its presence

- **Prognosis:**
  - Disease has not occurred and we want to know who is most likely to develop the disease

Both are amenable to multivariable approaches and prediction models

They are often mixed up

- Sometimes a diagnostic test itself can be used to predict future outcomes (e.g. PSA, Apgar)
  - E.g. With IGRA, we were hoping that they would be accurate for detecting LTBI as well as predicting who will develop TB disease
Types of diagnostic study designs (Phased approach)
Phases in intervention/drug trials

**Phase I:** Researchers test a new drug or treatment in a small group of people for the first time to evaluate its safety, determine a safe dosage range, and identify side effects.

**Phase II:** The drug or treatment is given to a larger group of people to see if it is effective and to further evaluate its safety.

**Phase III:** The drug or treatment is given to large groups of people to confirm its effectiveness, monitor side effects, compare it to commonly used treatments, and collect information that will allow the drug or treatment to be used safely.

**Phase IV:** Studies are done after the drug or treatment has been marketed to gather information on the drug’s effect in various populations and any side effects associated with long-term use.

---

Evidence base of clinical diagnosis

The architecture of diagnostic research

D L Sackett, R B Haynes

Considerable effort has been expended at the interface between clinical medicine and scientific methods to achieve the maximum validity and usefulness of diagnostic tests. This article focuses on the specific kinds of questions that arise in diagnostic research and the study architectures (the conversions of these clinical questions into appropriate research designs) used to answer them. As an example we shall take shall take assessment of the value of the plasma concentration of B-type natriuretic peptide (BNP) in the diagnosis of left ventricular dysfunction. Randomized controlled trials are dealt with elsewhere.

As in other forms of clinical research, there are several ways studying the potential or real diagnostic value of a physical sign or laboratory test, and each is appropriate to one kind of question and inappropriate for others. Among the possible questions about the relation between a putative diagnostic test and a target disorder (for example, the concentration of BNP and left ventricular dysfunction), four are most relevant.

Types of question

**Phase I questions**

Do test results in patients with the target disorder differ from those in normal people? Table 1 shows the architecture of this question.

For example, investigators at a British university hospital measured concentrations of BNP in non-systemic (“convenience”) samples from normal controls and from patients who had various cardiac conditions.

**Summary points**

Diagnostic studies should match methods to diagnostic questions

- Do test results in affected patients differ from those in normal individuals?
- Are patients with certain test results more likely to have the target disorder?
- Do test results distinguish patients with and without the target disorder among those in whom it is clinically sensible to suspect the disorder?
- Do patients undergoing the diagnostic test fare better than similar untreated patients?

The keys to validity in diagnostic test studies are independent, blind comparison of test results with a reference standard among a consecutive series of patients suspected (but not known) to have the target disorder.

Inclusion of missing and indeterminate results.

Replication of studies in other settings.

Both specificity and sensitivity may change as the same diagnostic test is applied in primary, secondary, and tertiary care.

BMJ 2002;324:539-41
Phase I to IV diagnostic studies

**Phase I questions**
- Do test results in patients with the target disorder differ from those in normal people?

*Table 1* Answering a phase I question: do patients with left ventricular dysfunction have higher concentrations of B-type natriuretic peptide (BNP) precursor than normal individuals?

<table>
<thead>
<tr>
<th>Patients known to have disorder</th>
<th>Normal controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median (range) concentration of BNP precursor (pg/ml)</td>
<td>493.5 (248.9-909.0)</td>
</tr>
</tbody>
</table>

*BMJ* 2002;324:539-41

**Phase II questions**
- Are patients with certain test results more likely to have the target disorder than patients with other test results?

*Table 2* Answering a phase II question: are patients with higher concentrations of B-type natriuretic peptide (BNP) more likely to have left ventricular dysfunction than patients with lower concentrations?

<table>
<thead>
<tr>
<th>Patients known to have target disorder</th>
<th>Normal controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>High BNP concentration</td>
<td>39</td>
</tr>
<tr>
<td>Normal BNP concentration</td>
<td>1</td>
</tr>
</tbody>
</table>

Test characteristics (95% CI):
- Sensitivity=88% (87% to 100%)
- Specificity=99% (77% to 98%)
- Positive predictive value=95% (84% to 99%)
- Negative predictive value=98% (81% to 100%)
- Likelihood ratio for an abnormal test result=13 (3.5 to 50.0)
- Likelihood ratio for a normal test result=0.03 (0.0003 to 0.19)

*BMJ* 2002;324:539-41
Phase I to IV diagnostic studies

**Phase III questions**
- Does the test result distinguish patients with and without the target disorder among patients in whom it is clinically reasonable to suspect that the disease is present?

Table 3 Answering a phase III question: among patients in whom it is clinically sensible to suspect left ventricular dysfunction (LVD), does the concentration of B-type natriuretic peptide (BNP) distinguish patients with and without left ventricular dysfunction?

<table>
<thead>
<tr>
<th>Concentration of BNP</th>
<th>Patients with LVD or echocardiography</th>
<th>Patients with normal results on echocardiography</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (&gt;10 pg/mL)</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Normal (&lt;10 pg/mL)</td>
<td>9</td>
<td>29</td>
</tr>
<tr>
<td>Prevalence (risk extent of LVD) 40/100=40%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test characteristics (95% CI):
- Sensitivity=88% (74% to 94%)
- Specificity=96% (92% to 99%)
- Positive predictive value=38% (29% to 48%)
- Negative predictive value=97% (90% to 99%)
- Likelihood ratio for an abnormal test result=1.3 (1.1 to 1.6)
- Likelihood ratio for a normal test result=0.4 (0.2 to 0.8)

*BMJ* 2002;324:539-41

**Phase IV questions**
- Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who are not tested?
Phased evaluation of medical tests

Levels/Phases
- Technical efficacy
- Intended use
- Diagnostic accuracy
- Usual range
- Subgroups
- Clinical population
- Diagnostic thinking efficacy
- Therapeutic efficacy
- Patient outcome efficacy
- Societal efficacy

Proposals for a Phased Evaluation of Medical Tests

Jeroen G. Lijnen, MD, PhD, Mariska Leeflang, PhD, Patrick M. M. Bossuyt, PhD

Med Decis Making 2009

RESEARCH METHODS & REPORTING

Assessing the value of diagnostic tests: a framework for designing and evaluating trials

The value of a diagnostic test is not simply measured by its accuracy, but depends on how it affects patient health. This article presents a framework for the design and interpretation of studies that evaluate the health consequences of new diagnostic tests

Lavinia Ferrante di Ruffano research fellow, Christopher J Hyde professor of public health and clinical epidemiology, Kristen J McCaffrey associate professor and principal research fellow, Patrick M M Bossuyt professor of clinical epidemiology, Jonathan J Deeks professor of biostatistics
Design is often decided by: what is the real or intended purpose of the test?

**TB examples**

- **Triage**: Urine LAM POC test in HIV+ to decide who needs further investigation for TB disease
- **Replacement**: Xpert MTB/RIF to replace sputum smear microscopy for investigating HIV+ TB suspects
- **Add-on**: IGRA added to TST for LTBI screening of HIV-infected persons with low CD4 counts

Most published TB Dx studies do not clearly indicate the intended purpose!
### TABLE 4. Guidelines on IGRA's: recommendations for HIV-infected populations

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Guideline or position statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>TST alone</td>
<td>WHO, Brazil</td>
</tr>
<tr>
<td>TST followed by IGRA, if TST positive</td>
<td>Spain</td>
</tr>
<tr>
<td>(and BCG-vaccinated)</td>
<td></td>
</tr>
<tr>
<td>TST followed by IGRA, if TST negative</td>
<td>Canada, Italy, Saudi Arabia, Spain, Ireland</td>
</tr>
<tr>
<td>Either TST or IGRA</td>
<td>Denmark, South Korea, Austria</td>
</tr>
<tr>
<td>Both TST and IGRA</td>
<td>ECDC, Portugal, Croatia, Slovakia, the Netherlands, USA (if either initial test negative), South Korea, UK</td>
</tr>
<tr>
<td>IGRA alone</td>
<td>Switzerland, Bulgaria, France, UK (if CD4 200–500)</td>
</tr>
<tr>
<td>No specific recommendations</td>
<td>Germany, Czech Republic, Norway, Japan, Finland, Australia</td>
</tr>
</tbody>
</table>

AAP, American Academy of Pediatrics; BCG, bacille Calmette–Guérin; CDC, US Centers for Disease Control and Prevention; ECDC, European Centre for Disease Prevention and Control; IGRA, interferon-gamma release assay; TST, tuberculin skin test; WHO, World Health Organization.

*Some countries/organizations are listed more than once because their recommendations vary across risk groups.*

---

**Fig. 1. Stepwise evaluation.**

- **New diagnostic test under evaluation**
  - **Always**
    - **Technique accuracy**
    - **Clinical pathway**
    - **Diagnostic accuracy**
    - **Patient outcome**
    - **Cost-effectiveness**
Replacement

- No change in consequences for TP, FP, FN, TN
- Accuracy may be enough (preferably paired data) – unless new test is more sensitive
- Other info needed: costs, safety, burden, indeterminate results...

Add on

- Potential change in consequences, also extra numbers (either extra positives or extra negatives)
- Extra testing: extra time, burden
- Other info needed: costs, safety, burden, indeterminate results...
- Effect of change in consequences (patient impact)
Triage

- May result in a completely different pathway and different population selected for treatment
- Accuracy will not be enough
- Other info needed: clinical impact, costs, safety, burden, indeterminate results...
- Advantage of early diagnosis?

A decision tree will be very helpful to clarify the intended purpose

Dowdy D et al.
When Is Measuring Sensitivity and Specificity Sufficient To Evaluate a Diagnostic Test, and When Do We Need Randomized Trials?

Sarah J. Lord, MD, MPH, MS; Les Irwig, MD, PhD; and R. John Simkins, MD, PhD

The clinical value of using a new diagnostic test depends on whether it improves patient outcomes beyond the outcomes achieved using an old diagnostic test. When can studies of diagnostic test accuracy provide sufficient information to infer clinical value, and when do clinicians need to wait for results from randomized trials? The authors argue that accuracy studies suffice if a new diagnostic test is equal or more specific than, but of similar sensitivity to, an old test. However, if a new test is more sensitive than an old test, it leads to the detection of extra cases of disease. Results from treatment trials that enrolled only patients detected by the old test may not apply to these extra cases. Clinicians need to wait for results from randomized trials assessing treatment efficacy in cases discovered by the new diagnostic test, unless they can be satisfied that the new test detects the same spectrum and subtype of disease as the old test or that treatment response is similar across the spectrum of disease.


For author affiliations, see end of text.

Figure 2: Assessing new tests using evidence of test accuracy, given that treatment is effective for cases detected by the old test.

1. New test vs. old test?
   - No
   - Yes

2. Is the new test more sensitive than the old test?
   - Yes
   - No

3. Do the extra cases detected respond to treatment?
   - Yes
   - No

4. Do the extra cases detected represent the same spectrum of disease as, given, and assuming, and the same outcomes as, disease detected by old test?
   - Yes
   - No

5. Are the new test have proper attributes, e.g., too unsafe, too specific, or not enough?
   - Yes
   - No

6. Is the new test diagnostic strategy known to be similar across the range of disease question or not?
   - Yes
   - No

7. Are the new test diagnostic strategy known to be effective or not?
   - Yes
   - No

8. Use new test.
   - No test or old test.

RCT = randomized controlled trial. * New test = diagnostic test that include the new test; old test = standard diagnostic strategies that do not include the new test.
Key issue to appreciate:

Accuracy may or may not result in clinical impact (on patient outcomes)
Rapid tests for influenza: Clinical impact

Impact of Rapid Diagnosis on Management of Adults Hospitalized With Influenza

“Impact” outcomes include:

- Change in clinical decisions
- Reduction in antibiotic use
- Increased antiviral use
- Decreased length of time to discharge
- Reduction in lab investigations, etc.

Most diagnostic studies are focused on technical and accuracy issues

Table 1. Hierarchy of Diagnostic Evaluation and the Number of Studies Available for Different Levels of Diagnostic Test in a Technology Assessment of Magnetic Resonance Spectroscopy for Brain Tumors

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Examples of Study Purpose or Measures</th>
<th>Studies Available, n</th>
<th>Patients, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Technical feasibility and optimization</td>
<td>Ability to produce consistent spectra</td>
<td>15</td>
<td>243</td>
</tr>
<tr>
<td>2</td>
<td>Diagnostic accuracy</td>
<td>Sensitivity and specificity</td>
<td>8</td>
<td>462</td>
</tr>
<tr>
<td>3</td>
<td>Diagnostic decision-making impact</td>
<td>Percentage of correct diagnosis, subjective assessment of diagnostic probabilities changed after the test</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>4</td>
<td>Therapeutic choice impact</td>
<td>Percentage of less therapy planned before SRS changed after the test</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>5</td>
<td>Patient outcome impact</td>
<td>Percentage of patients who improved with SRS compared with those without SRS (e.g., survival, quality of life)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Societal impact</td>
<td>Cost-effectiveness analysis (e.g., use of SRS to detect tumor in symptomatic population)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

* SRS = magnetic resonance spectroscopy.
Most existing tools and instruments are focused on test accuracy

Example:
- DEEP guidelines by TDR
- QUADAS tool
- STARD for better reporting
- Cochrane Handbook for Diagnostic Reviews
Main findings

- About 15% of all TB papers were mainly focused on TB diagnosis.
- Of these, about 85% were evaluation studies of tests and markers.
- Of these evaluation studies, about 85% are early phase studies of test accuracy; there are very little data on impact on patient outcomes.
- Most test accuracy studies are of moderate to low quality and are poorly reported.
- Essential methodological and design elements are often either not reported or poorly reported.
- These results have important implications for policy making.
GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies

The GRADE system can be used to grade the quality of evidence and strength of recommendations for diagnostic tests or strategies. This article explains how patient-important outcomes are taken into account in this process.

SUMMARY POINTS

As for other interventions, the GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic tests or strategies provides a comprehensive and transparent approach for developing recommendations. Cross sectional or cohort studies can provide high quality evidence of test accuracy.

However, test accuracy is a surrogate for patient-important outcomes, so such studies often provide low quality evidence for recommendations about diagnostic tests, even when the studies do not have serious limitations. Inferring from data on accuracy that a diagnostic test or strategy improves patient-important outcomes will require the availability of effective treatment, reduction of test-related adverse effects or anxiety, or improvement of patients’ wellbeing from prognostic information. Judgments are thus needed to assess the directness of test results in relation to consequences of diagnostic recommendations that are important to patients.

GRADE expectations are met in other fields that are well ahead of TB...

- Example: Rapid diagnostics tests (RIDTs) for influenza
  - 159 accuracy studies
  - 20+ impact studies (including several diagnostic RCTs)
In TB, since we have mostly accuracy data:
example from WHO EGM on tests for drug-resistant TB

<table>
<thead>
<tr>
<th>Test</th>
<th>n Studies (participants)</th>
<th>Design</th>
<th>Limitations</th>
<th>Directness</th>
<th>Inconsistency</th>
<th>Imprecise or sparse data</th>
<th>Publication Bias</th>
<th>Evidence Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODS</td>
<td>9 (1474)</td>
<td>CS &amp; CC</td>
<td>Low</td>
<td>No evidence</td>
<td>Low</td>
<td>Low</td>
<td>Possible</td>
<td>Moderate</td>
</tr>
<tr>
<td>NRA</td>
<td>19 (2304)</td>
<td>CS &amp; CC</td>
<td>Low</td>
<td>No evidence</td>
<td>Low</td>
<td>Low</td>
<td>Possible</td>
<td>Moderate</td>
</tr>
<tr>
<td>CRI</td>
<td>31 (2498)</td>
<td>CS &amp; CC</td>
<td>Low</td>
<td>No evidence</td>
<td>Low</td>
<td>Low</td>
<td>Possible</td>
<td>Moderate</td>
</tr>
<tr>
<td>TLA</td>
<td>3 (439)</td>
<td>CS &amp; CC</td>
<td>Low</td>
<td>No evidence</td>
<td>Low</td>
<td>High</td>
<td>Possible</td>
<td>Low</td>
</tr>
<tr>
<td>Phage</td>
<td>12 (2935)</td>
<td>CS &amp; CC</td>
<td>Moderate/High</td>
<td>No evidence</td>
<td>Moderate/High</td>
<td>Low</td>
<td>Probable</td>
<td>Very low</td>
</tr>
<tr>
<td>LPA</td>
<td>12 (4937)</td>
<td>CS &amp; CC</td>
<td>Low</td>
<td>No evidence</td>
<td>Low</td>
<td>Low</td>
<td>Possible</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Regardless of study quality, precision, consistency ... accuracy studies will never lead to High Quality Evidence

There are 60+ systematic reviews on TB tests, but almost all focus on sensitivity and specificity (accuracy)
Conclusions

- Test accuracy studies need to be done better and reported better.
- Need to go beyond test accuracy and generate evidence on:
  - Impact of test on patient important outcomes
  - Impact of test on diagnostic thinking and decision making
  - Incremental or added value beyond what is already in place
  - Time to diagnosis and treatment
  - Cost-effectiveness

Optimism bias in TB diagnostic research

Madhukar Pai, MD, PhD [madhukar.pai@mcgill.ca]
Jessica Minion, MD
McGill University, Montreal
While almost all trials with “positive” results on antidepressants had been published, trials with “negative” results submitted to the US Food and Drug Administration, with few exceptions, remained either unpublished or were published with the results presented so that they would appear “positive.”

Non-replicated studies and publication bias – especially in genetic and biomarker studies
TB diagnostic studies can be optimistic because of:

- Case-control studies
- Inappropriate comparison groups
- Insufficient validation in high TB/HIV burden settings
- Inappropriate data analytic methods and exclusions
- Industry-led studies that are not independently validated
- Optimistic package inserts based on mostly in-house studies
- Controlled studies by test developers that are not replicable in the real world
- Biomarkers that fail to get converted into good products
Many TB dx studies are case-control

A large % of TB serology studies were case-control studies

Confirmed TB cases Vs. Healthy controls (often from low-incidence countries)

<table>
<thead>
<tr>
<th>TABLE 3. Characteristics of study quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
</tr>
<tr>
<td>Study design</td>
</tr>
<tr>
<td>Cross-sectional</td>
</tr>
<tr>
<td>Case-control</td>
</tr>
<tr>
<td>Nested within observational study</td>
</tr>
<tr>
<td>Recruitment of participants</td>
</tr>
<tr>
<td>Consecutive or random</td>
</tr>
<tr>
<td>Consecutive or not reported</td>
</tr>
<tr>
<td>Criteria clearly described</td>
</tr>
<tr>
<td>Complete verification by use of the reference standard</td>
</tr>
<tr>
<td>Execution of test described in sufficient detail</td>
</tr>
<tr>
<td>Index test results blinded to reference standard?</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

* The description of the test execution was deemed insufficient in one study.

Steingart KR et al.

Spectrum bias (a form of selection bias)

- Extreme contrast
  - Case-control design
- Normal contrast (Indicated population)
  - Consecutively recruited patients in whom the disease is suspected

Extreme contrast (spectrum bias) can result in overestimation of test accuracy
We find this in TB as well:  
Example: PCR tests for TB meningitis

Case-control design results in optimistic accuracy

Case-control studies had a two-fold higher diagnostic odds ratios than cross-sectional studies

Table 4. Stratified analyses for the evaluation of heterogeneity among studies with in-house tests

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Number of studies</th>
<th>Summary diagnostic odds ratio* (95% CI)</th>
<th>Test for heterogeneity p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>Case-control</td>
<td>19</td>
<td>46.0 (9.0-190.2)</td>
</tr>
<tr>
<td></td>
<td>Cross-sectional</td>
<td>17</td>
<td>43.3 (2.2-83.3)</td>
</tr>
<tr>
<td></td>
<td>Blinded interpretation of test and/or reference standard results</td>
<td>Yes</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>17</td>
<td>42.5 (9.9-173.2)</td>
</tr>
<tr>
<td></td>
<td>Consecutive or random sampling of participants</td>
<td>Yes</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>17</td>
<td>46.6 (13.6-160.6)</td>
</tr>
<tr>
<td></td>
<td>Prospective data collection</td>
<td>Yes</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>17</td>
<td>55.8 (20.8-131.6)</td>
</tr>
</tbody>
</table>

*Random effects model. *x*2 test for heterogeneity. CI=confidence interval.
It is not uncommon to see TB test evaluations where:

- Cases come from a high-incidence country and controls from a low-incidence country
- Tests work well in a low-incidence country and fall apart in a high-incidence country
- Tests that work well in immunocompetent persons fail in populations with high HIV prevalence

Lack of discrimination in TB endemic settings: example

![Graph showing optical density (OD) values obtained from TB patients with active tuberculosis disease who resided in northern Tanzania (TZ-TB) and 32 healthy, bacille Calmette-Guérin (BCG)-vaccinated, Danish residents. The mean OD ± 1 SD is shown as horizontal lines.](image)

*Figure 3*: Dotplot showing the optical density (OD) values obtained from TB patients with active tuberculosis disease who resided in northern Tanzania (TZ-TB) and 32 healthy, BCG-vaccinated Danish residents. The mean OD ± 1 SD is shown as horizontal lines. The dotted line indicates the cutoff value, calculated as the mean OD ± 3 SDs for the 32 healthy Danish residents.
Variation in performance in high vs low endemic countries: example

T-cell interferon-γ release assays for the rapid immunodiagnosis of tuberculosis: clinical utility in high-burden vs. low-burden settings

Keertan Dheda, Richard van Zyl Smit, Motasim Badri, and Madhukar Pai

<table>
<thead>
<tr>
<th>Sensitivity (95% CI)</th>
<th>High incidence countries</th>
<th>Low incidence countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.65 (0.57–0.72)</td>
<td>Tisekwa et al. 2008</td>
<td>0.58 (0.46–0.70)</td>
</tr>
<tr>
<td>0.67 (0.59–0.74)</td>
<td>Pan et al. 2007</td>
<td>0.57 (0.47–0.66)</td>
</tr>
<tr>
<td>0.64 (0.58–0.70)</td>
<td>Adeleye et al. 2007</td>
<td>0.59 (0.52–0.66)</td>
</tr>
<tr>
<td>0.71 (0.65–0.77)</td>
<td>Raluy et al. 2008</td>
<td>0.65 (0.55–0.76)</td>
</tr>
</tbody>
</table>

Pooled sensitivity = 0.61 (0.54–0.68)

Chi-square = 3.14, df = 4 (P = 0.3663)

Inconsistency (I²) = 33.3%

HIV can prove to be the acid test for any test!
Example of MycoDot

MycoDot was hailed to be a breakthrough because it was a simple dipstick test.

Commercialized and marketed by Mossman Associates (with support of PATH)

Package insert: sensitivity of 70% and specificity of 95%
But when the test was evaluated in countries with high HIV prevalence, the performance was disastrous.

Sens in HIV+ = 26%
Sens in HIV+ = 25%

Despite these results, the test is still available on the market!

Analysis of diagnostic studies

- It is not uncommon to see researchers:
  - Excluding patients or controls with no definitive diagnoses
  - Excluding indeterminate or inconclusive results
  - Perform post-hoc “discrepant” analysis to move numbers within 2 x 2 tables
- Such analyses often result in spuriously inflated accuracy estimates
Example: exclusion of indeterminates can inflate accuracy estimates

**Role of Interferon Gamma Release Assay in Active TB Diagnosis among HIV Infected Individuals**

Babu Sudan Subhakar Khobor, Rajeshwar Shrikhande, Surekha Swaminathan, Venkatesam Pillai, Pashupati Parameshwaran, Manohar Raja

Abstract

Background: A rapid and specific test is urgently needed for tuberculosis (TB) diagnosis especially among human immunodeficiency virus (HIV) infected individuals. In this study, we assessed the sensitivity of interferon gamma release assay (IGRA) in active tuberculosis patients who were positive for HIV infection and compared it with that of tuberculin skin test (TST).

Methods/Principal Findings: A total of 105 HIV-TB patients who were made for anti-tuberculosis and anti-retroviral therapy were included for this study. Out of which, 88.9% were culture positive for M. Tuberculosis Quantiferon TB Gold Interferon Gamma Release Assay (IGRA) and TST positive. Out of 105 patients, 29.9% were smear positive. Of 105 patients, 84.2% were smear negative. IGRA was found positive in 71% of patients, while TST was positive in 99.1%. Out of 105 patients, 23.8% were culture negative. Out of 23 patients, 8.7% were culture positive. Out of 105 patients, 29.9% were smear positive. Of 29 patients, 8.7% were smear positive. Out of 105 patients, 7.6% were treatment failures. Out of 76 patients, 8.7% were treatment failures. Out of 105 patients, 7.6% were smear positive. Of 76 patients, 7.6% were smear positive. Out of 76 patients, 13.3% were treatment failures. Out of 105 patients, 11.3% were treatment failures. Out of 105 patients, 11.3% were treatment failures.

Conclusions: Our study shows that IGRA is a better test in comparison with TST. IGRA is useful in the diagnosis of active TB disease. IGRA has a better sensitivity and specificity in comparison with TST. IGRA is a better test in comparison with TST. IGRA is a better test in comparison with TST. IGRA is a better test in comparison with TST. IGRA is a better test in comparison with TST.

**Discrepancy Analysis: A Biased and an Unscientific Method for Estimating Test Sensitivity and Specificity**

Ahsan Hafeez

Center for Disease Control and Prevention, Division of STD Prevention, Atlanta, Georgia

ABSTRACT: Discrepancy analysis is a widely used technique for estimating test performance indices (sensitivity, specificity, etc.) of DNA amplification tests for detecting infectious diseases. It has recently been claimed that the discrepancy analysis-based estimates of specificity are typically biased due to those based on culture and that the discrepancy analysis-based specificity of higher appreciable bias. In this article, I show that these conclusions are incorrect. Using a typical example from the published literature, I show that the discrepancy analysis-based estimates of sensitivity and specificity can generate a significant and clinically important overestimation of the true sensitivity and specificity values. Moreover, I demonstrate that the concept of discrepancy analysis is profoundly flawed and unscientific. It violates a fundamental principle of diagnostic testing—the principle that the test should not be used to determine the true disease status. Thus, the major problem with discrepancy analysis is not only that it is biased but that it is unscientific. Therefore, discrepancy analysis should not be adopted for the evaluation of any diagnostic test, especially in the early stages of development of new tests.

KEYWORDS: Discrepancy analysis, sensitivity, specificity, DNA amplification tests, QIAamp DNA MicroKit
Industry involvement in drug trials and its impact on study outcomes and conclusions

Industry involvement in diagnostic studies?

Quality and Reporting of Diagnostic Accuracy Studies in TB, HIV and Malaria: Evaluation Using QUADAS and STARD Standards


About 40% of TB, HIV, Malaria diagnostic studies had industry involvement or known conflict of interest

Table 2. Characteristics of the studies included (N = 90)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td></td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>45 (50)</td>
</tr>
<tr>
<td>Malaria</td>
<td>10 (10)</td>
</tr>
<tr>
<td>HIV</td>
<td>27 (29)</td>
</tr>
<tr>
<td>Studying sample</td>
<td></td>
</tr>
<tr>
<td>Affix</td>
<td>10</td>
</tr>
<tr>
<td>Asia</td>
<td>21</td>
</tr>
<tr>
<td>Australia and Oceania</td>
<td>63</td>
</tr>
<tr>
<td>Europe</td>
<td>37</td>
</tr>
<tr>
<td>North America</td>
<td>13</td>
</tr>
<tr>
<td>South America</td>
<td>66</td>
</tr>
<tr>
<td>Number of patients per study</td>
<td>3.1 (1.6, 6.1)</td>
</tr>
<tr>
<td>Number of studies with industry involvement</td>
<td>22 (44)</td>
</tr>
<tr>
<td>Number of studies with conflict of interest</td>
<td>31 (42)</td>
</tr>
<tr>
<td>Year of publication</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>47 (71)</td>
</tr>
<tr>
<td>2005</td>
<td>23 (30)</td>
</tr>
<tr>
<td>2006</td>
<td>27 (30)</td>
</tr>
<tr>
<td>Number of journals where included studies were published</td>
<td>49</td>
</tr>
</tbody>
</table>
FASTPlaque tests for drug-resistant TB

Package inserts are always optimistic, but based on small in-house studies

Sensitivity = 98%
Specificity = 100%

Sensitivity = 93%
Specificity = 100%

Sensitivity = 100%
Specificity = 100%

Sensitivity = 100%
Specificity = 99%
### Test Results

<table>
<thead>
<tr>
<th>Test</th>
<th>Package insert sens</th>
<th>Package insert spec</th>
<th>Meta-analysis sens</th>
<th>Meta-analysis spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFT-Gold</td>
<td>89%</td>
<td></td>
<td>79%</td>
<td></td>
</tr>
<tr>
<td>FASTPlaque-Response</td>
<td>96 - 100%</td>
<td>99 - 100%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>Anda-TB IgG</td>
<td>85 - 90%</td>
<td>85 - 100%</td>
<td>60 - 75%</td>
<td>~90%</td>
</tr>
<tr>
<td>MycoDot</td>
<td>70%</td>
<td>95%</td>
<td>26% - 76%</td>
<td>84% - 97%</td>
</tr>
<tr>
<td>Clearview TB ELISA</td>
<td>81% (HIV+)</td>
<td>93 - 98%</td>
<td>56% (HIV+)</td>
<td>95%</td>
</tr>
<tr>
<td>GenoType MDTBDplus</td>
<td>99%</td>
<td>99%</td>
<td>98%</td>
<td>99%</td>
</tr>
<tr>
<td>Gen-Probe MTD</td>
<td>97% (S+)</td>
<td>100% (S+)</td>
<td>97% (S+)</td>
<td>96% (S+)</td>
</tr>
<tr>
<td></td>
<td>72% (S-)</td>
<td>99% (S-)</td>
<td>76% (S-)</td>
<td>95% (S-)</td>
</tr>
</tbody>
</table>

### Summary

**MODS:**
- **Sensitivity:** developed in Peru – performs excellent
  - Sensitivity better than LJ (98 vs. 84%)
  - Fast turnaround time (1 week vs. 6 weeks+)

**Issues:**
- Sensitivity 80%
- Issues with contamination
- Issues with reliability

---

**Microscopic-Observation Drug-Susceptibility Assay for the Diagnosis of TB**

David A. M. Morris, M.D., Csilla A. W. Evans, M.D., Ph.D., Robert H. Green, M.D.,
Louie Valenzuela, M.D., Jorge Gouveia, M.D., Abdel-Mageed, M.D., Efraín Sánchez, M.D.,
Yvette Falcón, M.D., Juan Carlos Sanz, M.D., Cayo Salazar, M.D.,
Richard Liburd, M.D., Marisela Jiménez, M.D.,
Daisy de la Cuesta, M.D., A. Robert & Escobar, M.D., Ph.D.,
and Sue R. Freedland, M.D., Ph.D.
FASTPlaque phage assay – performed well when done by industry

100% sens
100% spec

Implemented in Kenya – performs poorly

Despite upgrading the lab:
Low accuracy (31% sens; 95% spec)
Issues with contamination (nearly have were not interpretable)

Initial positive results that do not work out:
MPB64 skin patch test (Sequella Inc.)

Early data in 1998:
Sensitivity: 98%
Specificity: 100%

In 2012, still not commercially available