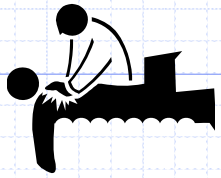
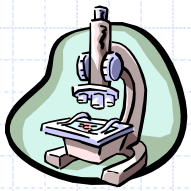
Diagnostic research designs: an introductory overview

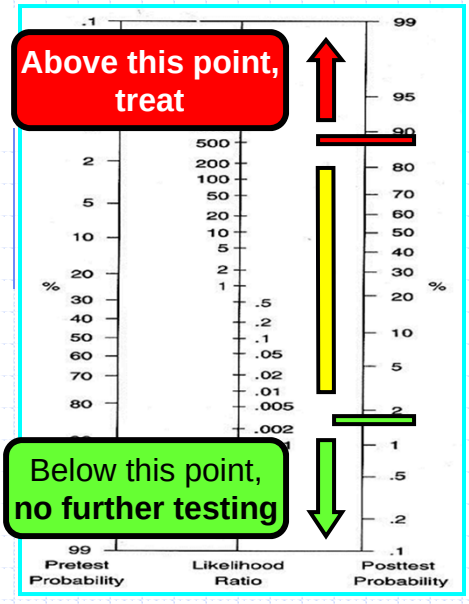




Madhukar Pai, MD, PhD
Associate Professor of Epidemiology, McGill University
Montreal, Canada

Email: madhukar.pai@mcgill.ca

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



Disease ruled IN

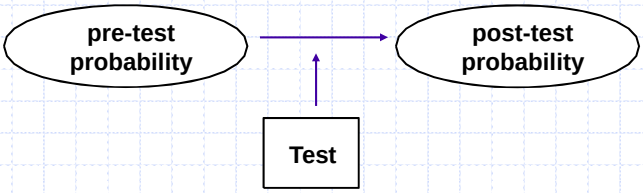
Disease not ruled in or out

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

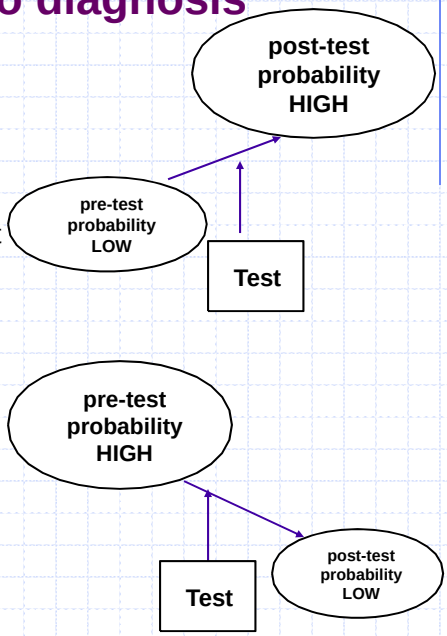
What you thought before + New information = What you think now



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).

Moons KGM. In: Grobbee & Hoes. Clinical Epidemiology. 2009

Some differences

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

Redundancy of Single Diagnostic Test Evaluation

Karel G.M. Moons,[1,2,3] Gerri-Anne van Es,[4] Bowine C. Michel,[5] Harry R. Büller,[6] J. Dik F. Habbema,[3] and Diederick E. Grobbee[1]

Moons et al. Epidemiology 1999

Diagnostic research

Diagnostic studies as multivariable, prediction research

K G M Moons, D E Grobbee

Patient outcomes in diagnostic research

Moons et al. JECH 2002

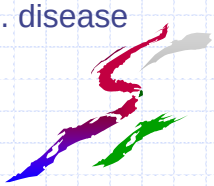
Opinion

Test Research versus Diagnostic Research

Moons et al. Clin Chem 2004

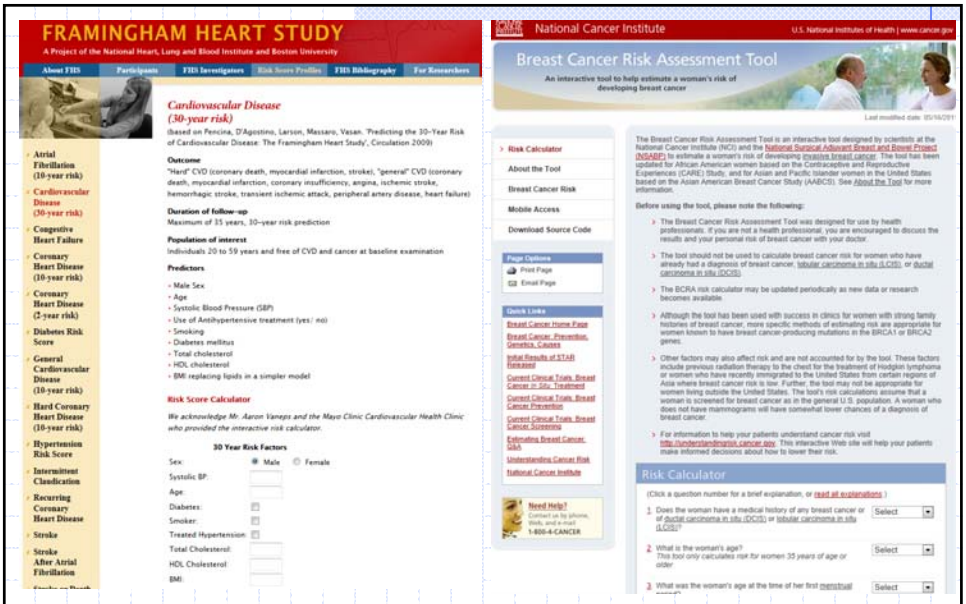
Diagnosis Vs Screening

- 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 IGRAs we were hoping that they will be accurate for detecting LTBI as well as predicting who will develop TB disease



PERSPECTIVE | Annals of Internal Medicine

Against Diagnosis

Andrew J. Vickers, PhD; Ethan Basch, MD; and Michael W. Kattan, PhD

The act of diagnosis requires that patients be placed in a binary category of either having or not having a certain disease. Accordingly, the diseases of particular concern for industrialized countries—such as type 2 diabetes, obesity, or depression—require that a somewhat arbitrary cut-point be chosen on a continuous scale of measurement (for example, a fasting glucose level >6.9 mmol/L [>125 mg/dL] for type 2 diabetes). These cut-points do not adequately reflect disease biology, may inappropriately treat patients on either side of the cut-point as 2 homogenous risk groups, fail to incorporate other risk factors, and are invariable to patient preference. This article discusses risk prediction as an alternative to diagnosis. Patient risk factors (blood pressure, age) are combined into a single statistical model (risk for a cardiovascular event within 10 years) and the results are used in shared decision making about possible treatments. The authors compare and contrast the diagnostic and risk prediction approaches and attempt to identify the types of medical problem to which each is best suited.

www.annals.org

Ann Intern Med. 2008;149:200-203.

For author affiliations, see end of text.

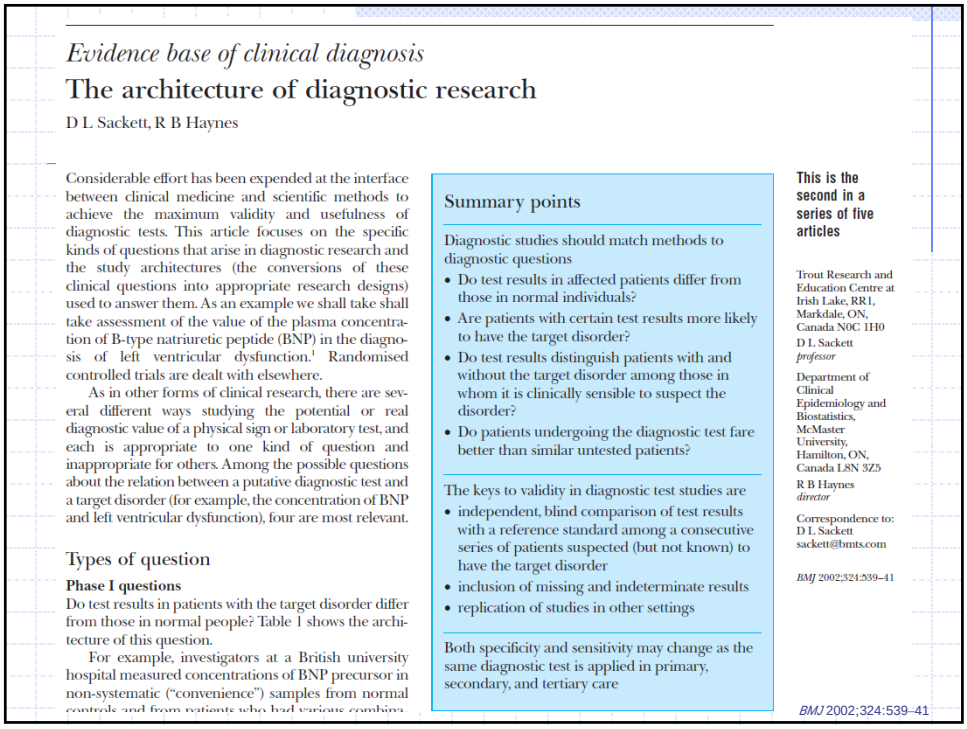
Table. Comparison of Typical Features of Diagnostic and Risk Prediction Approaches

Variable	Diagnosis	Risk Prediction
Approach	Patients are given a diagnosis: Either they have the disease or they do not	Patients are given a probability of a future event
Example	Syphilitic hepatitis	Cardiovascular event within 10 years
Lesion	Unambiguous	Nonexistent or equivocal
Example	Torn aorta	Depression
Treatment effectiveness	Often highly effective	Helpful, but patients may have event with treatment or avoid the event even if untreated
Example	Antibiotics for syphilis	Statins for high cholesterol level
Course of treatment	Dictated by diagnosis	Open to discussion
Example	Surgical treatment of a torn aorta	Treatment of early-stage prostate cancer
Patient preference	Generally of minor importance	Often of major importance
Example	Antibiotics for syphilis	Treatment of early-stage prostate cancer
Symptoms	Patient has distressing symptoms	Patient is often asymptomatic: Disorder is a risk factor for a future event
Example	Syphilitic hepatitis	Hyperlipidemia

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.[1] Randomised controlled trials are dealt with elsewhere.

As in other forms of clinical research, there are several different 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 precursor in non-systematic ("convenience") samples from normal controls and from patients who had various combina-

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 untested 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

This is the second in a series of five articles

Trout Research and Education Centre at Irish Lake, RR1, Markdale, ON, Canada N0C 1H0
D L Sackett
professor

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada L8N 3Z5
R B Haynes
director

Correspondence to:
D L Sackett
sackett@bmts.com

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?

	Patients known to have disorder	Normal controls
Median (range) concentration of BNP precursor (pg/ml)	493.5 (248.9-909.0)	129.4 (53.6-159.7)

BMJ 2002;324:539–41

Phase I to IV diagnostic studies

- **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?

	Patients known to have target disorder	Normal controls
High BNP concentration	39	2
Normal BNP concentration	1	25

Test characteristics (95% CI):
Sensitivity=98% (87% to 100%)
Specificity=92% (77% to 98%)
Positive predictive value=95% (84% to 99%)
Negative predictive value=96% (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?

	Patients with LVD on echocardiography	Patients with normal results on echocardiography
Concentration of BNP:		
High (>17.9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29
Prevalence (pretest probability) of LVD	40/126=32%	

Test characteristics (95% CI):
Sensitivity=88% (74% to 94%)
Specificity=34% (25% to 44%)
Positive predictive value=38% (29% to 48%)
Negative predictive value=85% (70% to 94%)
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 0.9)

BMJ 2002;324:539–41

Phase I to IV diagnostic studies

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

BMJ 2002;324:539–41

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. Lijmer, MD, PhD, Mariska Leeflang, PhD, Patrick M. M. Bossuyt, PhD

Med Desic Making 2009

BMJ

BMJ 2012;344:e686 doi: 10.1136/bmj.e686 (Published 21 February 2012)

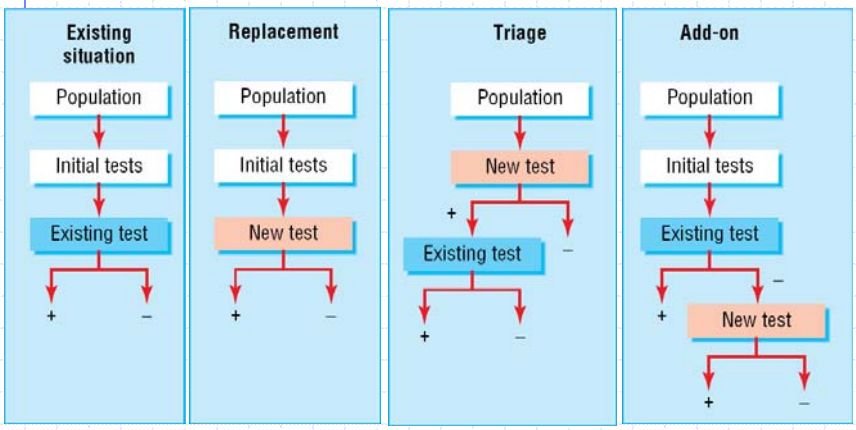
RESEARCH METHODS & REPORTING

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*[1], Christopher J Hyde *professor of public health and clinical epidemiology*[2], Kirsten J McCaffery *associate professor and principal research fellow*[3], Patrick M M Bossuyt *professor of clinical epidemiology*[4], Jonathan J Deeks *professor of biostatistics*[1]

Design is often decided by: what is the real or intended purpose of the test?



Bossuyt, BMJ, 2006

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 IGRAs: recommendations for HIV-infected populations

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

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

Denkinger C et al. Clin Micro Infect 2011

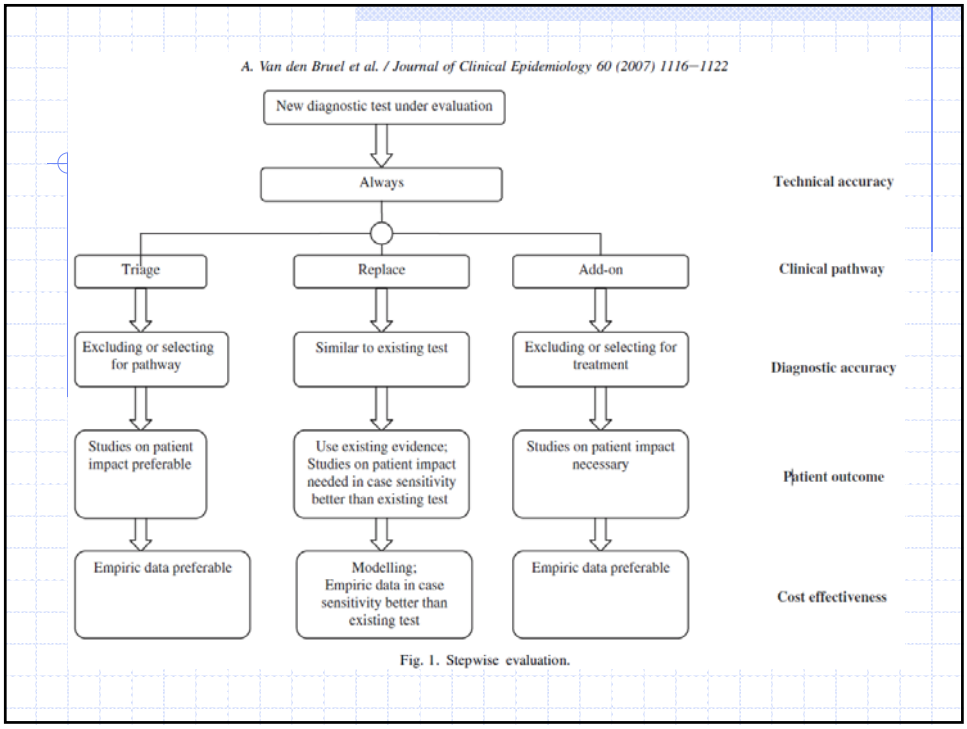
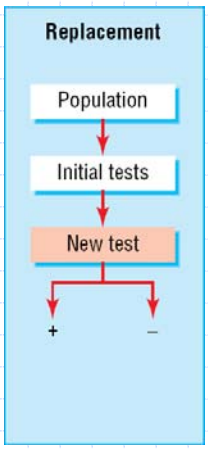


Fig. 1. Stepwise evaluation.

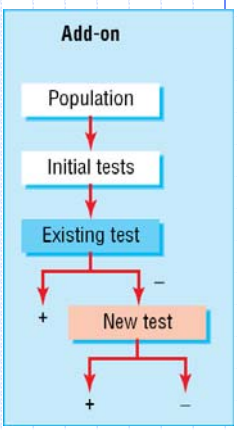
Replacement

- 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...



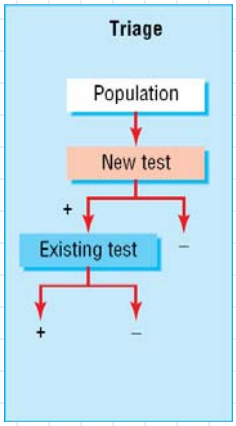
Add on

- 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)

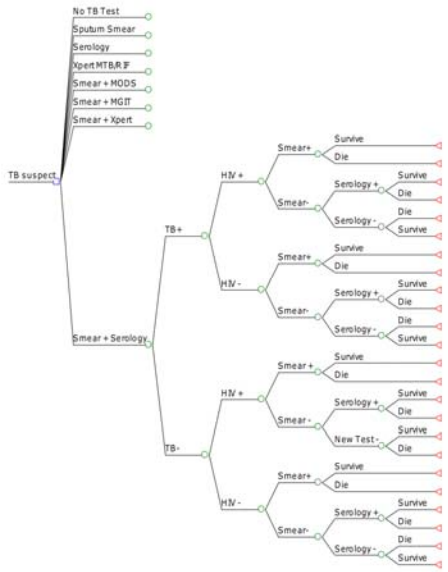


Triage

- 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.

Annals of Internal Medicine | ACADEMIA AND CLINIC

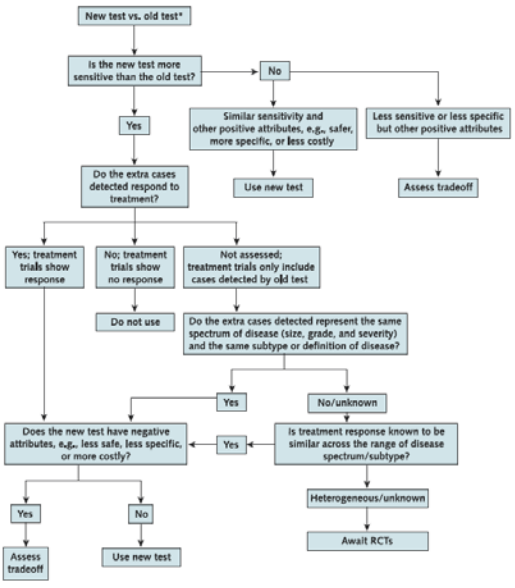
When Is Measuring Sensitivity and Specificity Sufficient To Evaluate a Diagnostic Test, and When Do We Need Randomized Trials?

Sarah J. Lord, MBBS, MS; Les Irwig, MBBCh, PhD; and R. John Simes, MBBS, MS, MD

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 safer 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 detected 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.

Ann Intern Med. 2006;144:850-855. www.annals.org
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.



RCT = randomized, controlled trial. * New test = diagnostic strategies 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)

Annals of Internal Medicine | ARTICLE

B-Type Natriuretic Peptide Testing, Clinical Outcomes, and Health Services Use in Emergency Department Patients With Dyspnea

A Randomized Trial

Hans-Gerhard Schneider, MBBS, MD; Louisa Lam, MPH; Amaali Lokuge, MBBS; Henry Krum, MBBS, PhD; Matthew T. Naughton, MBBS; Pieter De Villiers Smit, MBBS; Adam Bystrzycki, MBBS; David Eccleston, MBBS, PhD; Jacob Federman, MBBS; Genevieve Flannery, MBBS; and Peter Cameron, MBBS, MD

Background: B-type natriuretic peptide (BNP) is used to diagnose heart failure, but the effects of using the test on all dyspneic patients is uncertain.

Objective: To assess whether BNP testing alters clinical outcomes and health services use of acutely dyspneic patients.

Design: Randomized, single-blind study. Patients were assigned to a treatment group through randomized numbers in a sealed envelope. Patients were blinded to the intervention, but clinicians and those who assessed trial outcomes were not.

Setting: 2 Australian teaching hospital emergency departments.

Patients: 612 consecutive patients who presented with acute severe dyspnea from August 2005 to March 2007.

Intervention: BNP testing (n = 306) or no testing (n = 306).

Measurements: Admission rates, length of stay, and emergency department medications (primary outcomes); mortality and readmission rates (secondary outcomes).

Results: There were no between-group differences in hospital admission rates (85.6% [BNP group] vs. 86.6% [control group]; difference, −1.0 percentage point [95% CI, −6.5 to 4.5 percentage points]; P = 0.73), length of admission (median, 4.4 days [interquartile range, 2 to 9 days] vs. 5.0 days [interquartile range, 2 to 9 days]; P = 0.94), or management of patients in the emergency department. Test discrimination was good (area under the receiver-operating characteristic curve, 0.87 [CI, 0.83 to 0.91]). Adverse events were not measured.

Limitation: Most patients were very short of breath and required hospitalization; the findings might not apply for evaluating patients with milder degrees of breathlessness.

Conclusion: Measurement of BNP in all emergency department patients with severe shortness of breath had no apparent effects on clinical outcomes or use of health services. The findings do not support routine use of BNP testing in all severely dyspneic patients in the emergency department.

Primary Funding Source: Janssen-Cilag.

Ann Intern Med. 2009;150:365-371. www.annals.org
For author affiliations, see end of text.
ClinicalTrials.gov registration number: NCT00163709.

Rapid tests for influenza: Clinical impact

Impact of the Rapid Diagnosis of Influenza on Physician Decision-Making and Patient Management in the Pediatric Emergency Department: Results of a Randomized, Prospective, Controlled Trial

Aleta B. Bonner, DVM, MD*; Kathy W. Monroe, MD*; Lynya I. Talley, PhDS; Ann E. Klasner, MD, MPH*; and David W. Kimberlin, MD‡

ABSTRACT. *Objective.* To determine the impact of the rapid diagnosis of influenza on physician decision-making and patient management, including laboratory tests and radiographs ordered, patient charges associated with these tests, antibiotics/antivirals prescribed, and length of time to patient discharge from the emergency department.

Methods. Patients aged 2 months to 21 years presenting to an urban children's teaching hospital emergency department were screened for fever and cough, coryza, myalgias, headache, and/or malaise. After obtaining informed consent, patients were randomized to 1 of 2 groups: 1) physician receives (physician aware of) the rapid influenza test result; or 2) physician does not receive (physician unaware of) the result. For patients in the physician aware group, nasopharyngeal swabs were obtained, immediately tested with the FluCHA test for influenza A and B, and the result was placed on the chart before patient evaluation by the attending physician. For the physician unaware group, nasopharyngeal swabs were obtained, stored according to manufacturer's directions, and tested within 24 hours. Results for the physician unaware group were not disclosed to the treating physicians at any time. The 2 resultant influenza-positive groups (aware and unaware) were compared for laboratory and radiograph studies and their associated patient charges, antibiotic/antiviral prescriptions, and length of stay in the emergency department.

Results. A total of 418 patients were enrolled, and 391 completed the study. Of these, 202 tested positive for influenza. Comparison of the 96 influenza-positive patients whose physician was aware of the result with the 106 influenza-positive patients whose physician was unaware of the result revealed significant reductions among the former group in: 1) numbers of complete blood counts, blood cultures, urinalyses, urine cultures, and chest radiographs performed; 2) charges associated with these tests; 3) antibiotics prescribed; and 4) length of stay in the emergency department. The number of influenza-positive patients who received prescriptions for antiviral drugs was significantly higher among those whose physician was aware of the result.

Conclusions. Physician awareness of a rapid diagnosis of influenza in the pediatric emergency department significantly reduced the number of laboratory tests and radiographs ordered and their associated charges, decreased antibiotic use, increased antiviral use, and decreased length of time to discharge. *Pediatrics* 2003;112:363–367; *pediatric, influenza, physician decision-making, patient management.*

Influenza virus types A and B are common respiratory pathogens in the pediatric population. Depending on age, attack rates may be 1.5 to 3 times higher than for adults, with school-aged children having the highest attack rates.[1,2] A retrospective cohort study of children under 15 years of age demonstrated outpatient visits attributable to influenza ranging from 6 to 15 per 100 children.[3] Infection with influenza virus leads to a significant increase in primary care visits, and also increases in emergency department utilization during wintertime epidemics.[3]

Rapid diagnostic test kits for influenza types A and B are currently available for outpatient use and have proven to be both sensitive and specific.[4–7] Few studies have been performed which analyze the impact of rapid diagnostic testing for influenza and subsequent effect on patient management.[8–11] To date, there are no prospective, randomized studies analyzing use of rapid influenza testing and effect on patient management in the pediatric emergency department. Rapid diagnostic tests are not currently routinely incorporated in the work-up of infants and children with fever and vague symptoms, or with fever and no documented source.[12] Use of rapid tests in the pediatric emergency department which are sensitive and specific for influenza could potentially decrease performance of other more invasive tests, thereby reducing associated patient charges, reducing patient length of stay in the emergency depart-

Pediatrics 2003;112;363-367

Impact of Rapid Diagnosis on Management of Adults Hospitalized With Influenza

ARCHIVES EXPRESS

Ann R. Falsey, MD; Yoshihiko Murata, MD, PhD; Edward E. Walsh, MD

Background: Rapid influenza testing decreases antibiotic and ancillary test use in febrile children, yet its effect on the care of hospitalized adults is unexplored. We compared the clinical management of patients with influenza whose rapid antigen test result was positive (Ag+) with the management of those whose rapid antigen test result was negative or the test was not performed (Ag0).

Methods: Medical record review was performed on patients with influenza hospitalized during 4 winters (1999–2003). Hospital policy mandated influenza testing (antigen or culture) for all patients with acute cardiopulmonary diseases admitted from November 15 through April 15. A subset of patients participated in an epidemiological study and had reverse-transcriptase polymerase chain reaction or serologic testing performed. Clinical data from Ag+ and Ag0 patients were compared.

Results: Of 166 patients with available records, 86 were Ag+ and 80 were Ag0. Antibiotic use (74 [86%] of 86 patients vs 79 [99%] of 80 patients; P=.002) was less and antibiotic discontinuance (12 [14%] of 80 patients vs 2 [2%] of 80 patients; P=.01) was greater in Ag+ compared with Ag0 patients. No significant differences in antibiotic days, length of hospital stay, or antibiotic complications were noted. Antiviral use (63 [73%] of 86 patients vs 6 [8%] of 80 patients; P<.001) was greater in Ag+ than Ag0 patients. Antigen status was independently associated with withholding or discontinuing antibiotics in multivariate analysis. Of 44 Ag+ patients deemed low risk for bacterial infection, 27 continued to receive antibiotics despite positive influenza test results. These patients more commonly had pulmonary disease and had significantly more abnormal lung examination results (P=.005) compared with those in whom antibiotics were withheld or discontinued.

Conclusions: Rapid influenza testing leads to reductions in antibiotic use in hospitalized adults. Better tools to rule out concomitant bacterial infection are needed to optimize the impact of viral testing.

Arch Intern Med. 2007;167:354-360

"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

33

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*

Level	Description	Examples of Study Purpose or Measures	Studies Available, *n*	Patients, *n*
1	Technical feasibility and optimization	Ability to produce consistent spectra	85	2434
2	Diagnostic accuracy	Sensitivity and specificity	8	461
3	Diagnostic thinking impact	Percentage of times clinicians' subjective assessment of diagnostic probabilities changed after the test	2	32
4	Therapeutic choice impact	Percentage of times therapy planned before MRS changed after the test	2	105
5	Patient outcome impact	Percentage of patients who improved with MRS diagnosis compared with those without MRS (e.g., survival, quality of life)	0	0
6	Societal impact	Cost-effectiveness analysis (e.g., use to detect tumor in asymptomatic population)	0	0

* MRS = magnetic resonance spectroscopy.

Tatsioni, Annals

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

Mapping the landscape and quality of TB diagnostic research

Madhukar Pai, MD, PhD
Laurence Brunet, MSc
Jessica Minion, MD
Karen Steingart, MD, MPH
Andrew Ramsay, MSc
Christian Lienhardt, MD, PhD

Contact: madhukar.pai@mcgill.ca

McGill · Stop TB Partnership · Francis J. Curry National Tuberculosis Center · TDR For research on diseases of poverty · World Health Organization

36

OPEN ACCESS Freely available online

PLoS one

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

Patricia Scolari Fontela[1], Nitika Pant Pai[2], Ian Schiller[2], Nandini Dendukuri[2], Andrew Ramsay[3], Madhukar Pai[1,4]*

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, 2 Department of Medicine, Division of Clinical Epidemiology, McGill University, Montreal, Canada, 3 Special Programme for Research and Training in Tropical Diseases, World Health Organization, Geneva, Switzerland, 4 Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, Montreal, Canada

Abstract

Background: Poor methodological quality and reporting are known concerns with diagnostic accuracy studies. In 2003, the QUADAS tool and the STARD standards were published for evaluating the quality and improving the reporting of diagnostic studies, respectively. However, it is unclear whether these tools have been applied to diagnostic studies of infectious diseases. We performed a systematic review on the methodological and reporting quality of diagnostic studies in TB, malaria and HIV.

Methods: We identified diagnostic accuracy studies of commercial tests for TB, malaria and HIV through a systematic search of the literature using PubMed and EMBASE (2004–2006). Original studies that reported sensitivity and specificity data were included. Two reviewers independently extracted data on study characteristics and diagnostic accuracy, and used QUADAS and STARD to evaluate the quality of methods and reporting, respectively.

Findings: Ninety (38%) of 238 articles met inclusion criteria. All studies had design deficiencies. Study quality indicators that were met in less than 25% of the studies included adequate description of withdrawals (6%) and reference test execution (10%), absence of index test review bias (19%) and reference test review bias (24%), and report of uninterpretable results (22%). In terms of quality of reporting, 9 STARD indicators were reported in less than 25% of the studies: methods for calculation and estimates of reproducibility (0%), adverse effects of the diagnostic tests (1%), estimates of diagnostic accuracy between subgroups (10%), distribution of severity of disease/other diagnoses (11%), number of eligible patients who did not participate in the study (14%), blinding of the test readers (16%), and description of the team executing the test and management of indeterminate/outlier results (both 17%). The use of STARD was not explicitly mentioned in any study. Only 22% of 46 journals that published the studies included in this review required authors to use STARD.

Conclusion: Recently published diagnostic accuracy studies on commercial tests for TB, malaria and HIV have moderate to low quality and are poorly reported. The more frequent use of tools such as QUADAS and STARD may be necessary to improve the methodological and reporting quality of future diagnostic accuracy studies in infectious diseases.

PLoS One 2009

37

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

38

ANALYSIS

RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

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

39

BMJ 2008

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)

40

In TB, since we have mostly accuracy data:

example from WHO EGM on tests for drug-resistant TB

Test, # Studies (participants)	Design	Limitations	Directness	Inconsistency	Imprecise or sparse data	Publication Bias	Evidence Quality
MODS, 9 (1474)	CS & CC	Low	No evidence -1	Low	Low	Possible	Moderate
NRA, 19 (2304)	CS & CC	Low	No evidence -1	Low	Low	Possible	Moderate
CRI, 31 (2498)	CS & CC	Low	No evidence -1	Low	Low	Possible	Moderate
TLA, 3 (439)	CS & CC	Low	No evidence -1	Low	High -1	Possible	Low
Phage, 12 (2935)	CS & CC	Moderate/High -1	No evidence -1	Moderate/High -1	Low	Probable	Very low
LPA, 12 (4937)	CS & CC	Low	No evidence -1	Low	Low	Possible	Moderate

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

41

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

42

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

43

Optimism bias in TB diagnostic research

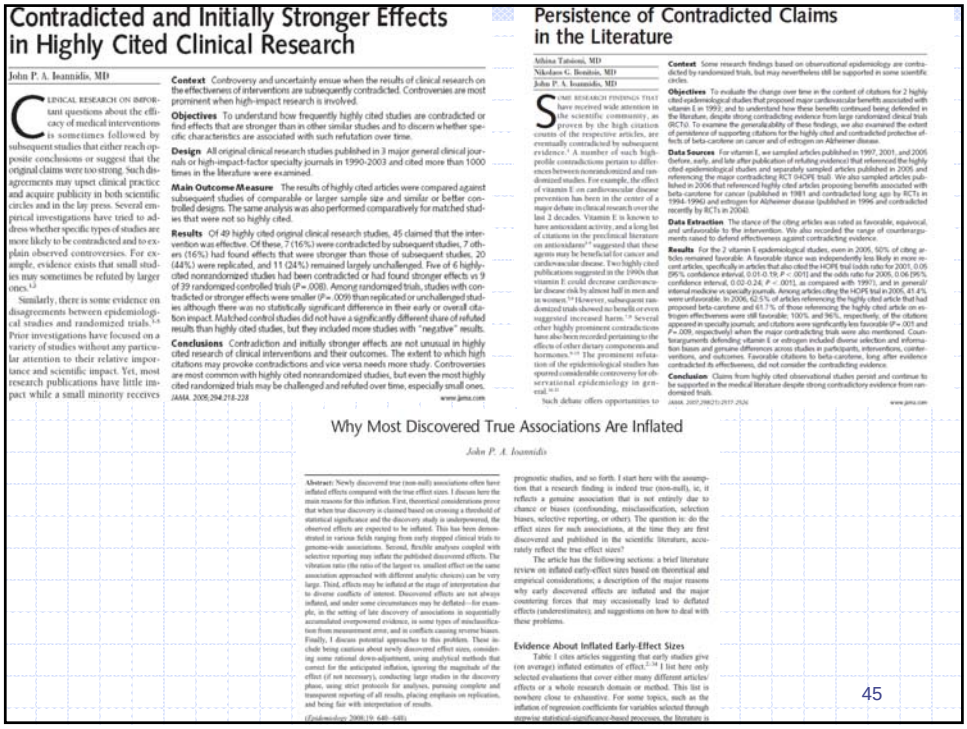
Madhukar Pai, MD, PhD [madhukar.pai@mcgill.ca]
Jessica Minion, MD

McGill University, Montreal

McGill

L'Institut de recherche du Centre universitaire de santé McGill
The Research Institute of the McGill University Health Centre
Les meilleurs soins pour la vie
The Best Care for Life

44



Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

CLINICAL RESEARCH ON IMPORTANT questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to address whether specific types of studies are more likely to be contradicted and to explain observed controversies. For example, evidence exists that small studies may sometimes be refuted by larger ones.[1,2]

Similarly, there is some evidence on disagreements between epidemiological studies and randomized trials.[3-5] Prior investigations have focused on a variety of studies without any particular attention to their relative importance and scientific impact. Yet, most research publications have little impact while a small minority receives

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discern whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.

Results Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials ($P = .008$). Among randomized trials, studies with contradicted or stronger effects were smaller ($P = .009$) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with "negative" results.

Conclusions Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. The extent to which high citations may provoke contradictions and vice versa needs more study. Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

JAMA. 2005;294:218-228

www.jama.com

Persistence of Contradicted Claims in the Literature

Athina Tatsioni, MD
Nikolaos G. Bonitsis, MD
John P. A. Ioannidis, MD

SOME RESEARCH FINDINGS THAT have received wide attention in the scientific community, as proven by the high citation counts of the respective articles, are eventually contradicted by subsequent evidence.[1] A number of such high-profile contradictions pertain to differences between nonrandomized and randomized studies. For example, the effect of vitamin E on cardiovascular disease prevention has been in the center of a major debate in clinical research over the last 2 decades. Vitamin E is known to have antioxidant activity, and a long list of citations in the preclinical literature on antioxidants[2,3] suggested that these agents may be beneficial for cancer and cardiovascular disease. Two highly cited publications suggested in the 1990s that vitamin E could decrease cardiovascular disease risk by almost half in men and in women.[4,5] However, subsequent randomized trials showed no benefit or even suggested increased harm.[7,8] Several other highly prominent contradictions have also been recorded pertaining to the effects of other dietary components and hormones.[9-11] The prominent refutation of the epidemiological studies has spurred considerable controversy for observational epidemiology in general.[6,12]

Such debate offers opportunities to

Context Some research findings based on observational epidemiology are contradicted by randomized trials, but may nevertheless still be supported in some scientific circles.

Objectives To evaluate the change over time in the content of citations for 2 highly cited epidemiological studies that proposed major cardiovascular benefits associated with vitamin E (in 1993) and to understand how these benefits continued being defended in the literature, despite strong contradicting evidence from large randomized clinical trials (RCTs). To examine the generalizability of these findings, we also examined the extent of persistence of supporting citations for the highly cited and contradicted protective effects of beta-carotene on cancer and of estrogen on Alzheimer disease.

Data Sources For vitamin E, we sampled articles published in 1997, 2001, and 2005 (before, early, and late after publication of refuting evidence) that referenced the highly cited epidemiological studies and separately sampled articles published in 2005 and referencing the major contradicting RCT (HOPE trial). We also sampled articles published in 2006 that referenced highly cited articles proposing benefits associated with beta-carotene for cancer (published in 1981 and contradicted long ago by RCTs in 1994-1996) and estrogen for Alzheimer disease (published in 1996 and contradicted recently by RCTs in 2004).

Data Extraction The stance of the citing articles was rated as favorable, equivocal, and unfavorable to the intervention. We also recorded the range of counterarguments raised to defend effectiveness against contradicting evidence.

Results For the 2 vitamin E epidemiological studies, even in 2005, 50% of citing articles remained favorable. A favorable stance was independently less likely in more recent articles, specifically in articles that also cited the HOPE trial (odds ratio for 2001, 0.06 [95% confidence interval, 0.01-0.19; $P < .001$] and the odds ratio for 2005, 0.06 [95% confidence interval, 0.02-0.24; $P < .001$], as compared with 1997), and in general/internal medicine vs specialty journals. Among articles citing the HOPE trial in 2005, 45.4% were unfavorable. In 2006, 62.5% of articles referencing the highly cited article that had proposed beta-carotene and 61.7% of those referencing the highly cited article on estrogen effectiveness were still favorable; 100% and 96%, respectively, of the citations appeared in specialty journals; and citations were significantly less favorable ($P < .001$ and $P = .009$, respectively) when the major contradicting trials were also mentioned. Counterarguments defending vitamin E or estrogen included diverse selection and information biases and genuine differences across studies in participants, interventions, contraventions, and outcomes. Favorable citations to beta-carotene, long after evidence contradicted its effectiveness, did not consider the contradicting evidence.

Conclusion Claims from highly cited observational studies persist and continue to be supported in the medical literature despite strong contradictory evidence from randomized trials.

JAMA. 2007;298(21):2517-2526

www.jama.com

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

Abstract: Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes. I discuss here the main reasons for this inflation. First, theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated. This has been demonstrated in various fields ranging from early stopped clinical trials to genome-wide associations. Second, flexible analyses coupled with selective reporting may inflate the published discovered effects. The vibration ratio (the ratio of the largest vs. smallest effect on the same association approached with different analytic choices) can be very large. Third, effects may be inflated at the stage of interpretation due to diverse conflicts of interest. Discovered effects are not always inflated, and under some circumstances may be deflated—for example, in the setting of late discovery of associations in sequentially accumulated overpowered evidence, in some types of misclassification from measurement error, and in conflicts causing reverse biases. Finally, I discuss potential approaches to this problem. These include being cautious about newly discovered effect sizes, considering some rational down-adjustment, using analytical methods that correct for the anticipated inflation, ignoring the magnitude of the effect (if not necessary), conducting large studies in the discovery phase, using strict protocols for analyses, pursuing complete and transparent reporting of all results, placing emphasis on replication, and being fair with interpretation of results.

(*Epidemiology*. 2008;19:640–648)

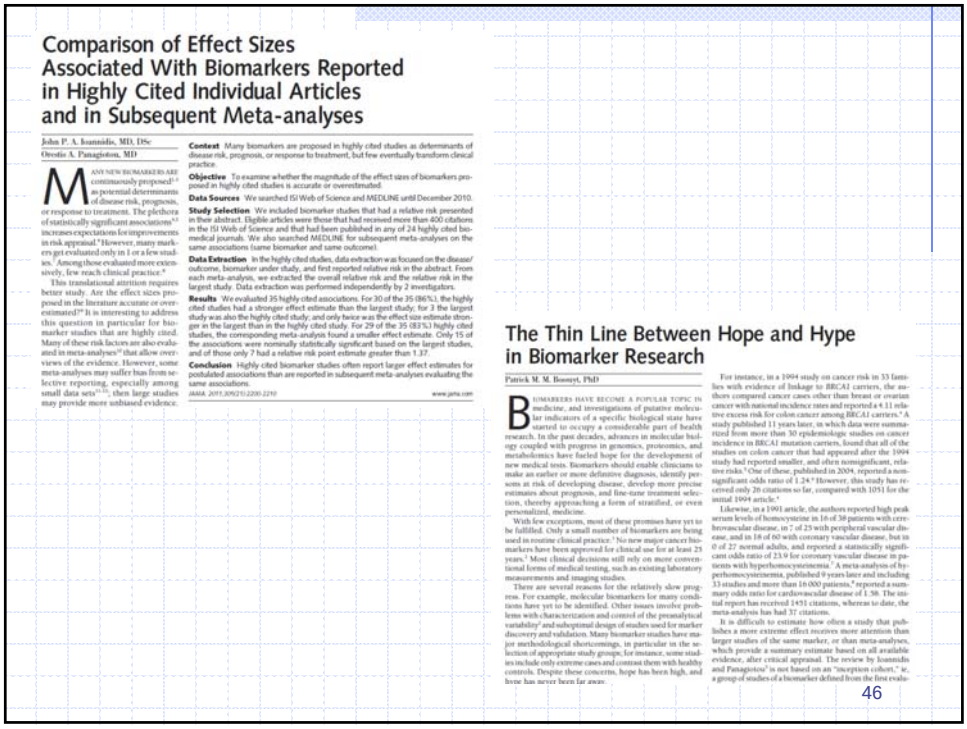
prognostic studies, and so forth. I start here with the assumption that a research finding is indeed true (non-null), ie, it reflects a genuine association that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other). The question is: do the effect sizes for such associations, at the time they are first discovered and published in the scientific literature, accurately reflect the true effect sizes?

The article has the following sections: a brief literature review on inflated early-effect sizes based on theoretical and empirical considerations; a description of the major reasons why early discovered effects are inflated and the major countering forces that may occasionally lead to deflated effects (underestimates); and suggestions on how to deal with these problems.

Evidence About Inflated Early-Effect Sizes

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.[7–38] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is

45



Comparison of Effect Sizes Associated With Biomarkers Reported in Highly Cited Individual Articles and in Subsequent Meta-analyses

John P. A. Ioannidis, MD, DSc
Orestis A. Panagiotou, MD

MANY NEW BIOMARKERS ARE continuously proposed[1-3] as potential determinants of disease risk, prognosis, or response to treatment. The plethora of statistically significant associations[4,5] increases expectations for improvements in risk appraisal.[6] However, many markers get evaluated only in 1 or a few studies.[7] Among those evaluated more extensively, few reach clinical practice.[8]

This translational attrition requires better study. Are the effect sizes proposed in the literature accurate or overestimated?[9] It is interesting to address this question in particular for biomarker studies that are highly cited. Many of these risk factors are also evaluated in meta-analyses[10] that allow overviews of the evidence. However, some meta-analyses may suffer bias from selective reporting, especially among small data sets[11,12]; then large studies may provide more unbiased evidence.

Context Many biomarkers are proposed in highly cited studies as determinants of disease risk, prognosis, or response to treatment, but few eventually transform clinical practice.

Objective To examine whether the magnitude of the effect sizes of biomarkers proposed in highly cited studies is accurate or overestimated.

Data Sources We searched ISI Web of Science and MEDLINE until December 2010.

Study Selection We included biomarker studies that had a relative risk presented in their abstract. Eligible articles were those that had received more than 400 citations in the ISI Web of Science and that had been published in any of 24 highly cited biomedical journals. We also searched MEDLINE for subsequent meta-analyses on the same associations (same biomarker and same outcome).

Data Extraction In the highly cited studies, data extraction was focused on the disease/outcome, biomarker under study, and first reported relative risk in the abstract. From each meta-analysis, we extracted the overall relative risk and the relative risk in the largest study. Data extraction was performed independently by 2 investigators.

Results We evaluated 35 highly cited associations. For 30 of the 35 (86%), the highly cited studies had a stronger effect estimate than the largest study; for 3 the largest study was also the highly cited study; and only twice was the effect size estimate stronger in the largest than in the highly cited study. For 29 of the 35 (83%) highly cited studies, the corresponding meta-analysis found a smaller effect estimate. Only 15 of the associations were nominally statistically significant based on the largest studies, and of those only 7 had a relative risk point estimate greater than 1.37.

Conclusion Highly cited biomarker studies often report larger effect estimates for postulated associations than are reported in subsequent meta-analyses evaluating the same associations.

JAMA. 2011;305(21):2200-2210

www.jama.com

The Thin Line Between Hope and Hype in Biomarker Research

Patrick M. M. Bossuyt, PhD

BIOMARKERS HAVE BECOME A POPULAR TOPIC IN medicine, and investigations of putative molecular indicators of a specific biological state have started to occupy a considerable part of health research. In the past decades, advances in molecular biology coupled with progress in genomics, proteomics, and metabolomics have fueled hope for the development of new medical tests. Biomarkers should enable clinicians to make an earlier or more definitive diagnosis, identify persons at risk of developing disease, develop more precise estimates about prognosis, and fine-tune treatment selection, thereby approaching a form of stratified, or even personalized, medicine.

With few exceptions, most of these promises have yet to be fulfilled. Only a small number of biomarkers are being used in routine clinical practice.[1] No new major cancer biomarkers have been approved for clinical use for at least 25 years.[2] Most clinical decisions still rely on more conventional forms of medical testing, such as existing laboratory measurements and imaging studies.

There are several reasons for the relatively slow progress. For example, molecular biomarkers for many conditions have yet to be identified. Other issues involve problems with characterization and control of the preanalytical variability[3] and suboptimal design of studies used for marker discovery and validation. Many biomarker studies have major methodological shortcomings, in particular in the selection of appropriate study groups; for instance, some studies include only extreme cases and contrast them with healthy controls. Despite these concerns, hope has been high, and hype has never been far away.

For instance, in a 1994 study on cancer risk in 33 families with evidence of linkage to *BRCA1* carriers, the authors compared cancer cases other than breast or ovarian cancer with national incidence rates and reported a 4.11 relative excess risk for colon cancer among *BRCA1* carriers.[4] A study published 11 years later, in which data were summarized from more than 30 epidemiologic studies on cancer incidence in *BRCA1* mutation carriers, found that all of the studies on colon cancer that had appeared after the 1994 study had reported smaller, and often nonsignificant, relative risks.[5] One of these, published in 2004, reported a nonsignificant odds ratio of 1.24.[6] However, this study has received only 26 citations so far, compared with 1051 for the initial 1994 article.[4]

Likewise, in a 1991 article, the authors reported high peak serum levels of homocysteine in 16 of 38 patients with cerebrovascular disease, in 7 of 25 with peripheral vascular disease, and in 18 of 60 with coronary vascular disease, but in 0 of 27 normal adults, and reported a statistically significant odds ratio of 23.9 for coronary vascular disease in patients with hyperhomocysteinemia.[7] A meta-analysis of hyperhomocysteinemia, published 9 years later and including 33 studies and more than 16 000 patients,[8] reported a summary odds ratio for cardiovascular disease of 1.38. The initial report has received 1451 citations, whereas to date, the meta-analysis has had 37 citations.

It is difficult to estimate how often a study that publishes a more extreme effect receives more attention than larger studies of the same marker, or than meta-analyses, which provide a summary estimate based on all available evidence, after critical appraisal. The review by Ioannidis and Panagiotou[1] is not based on an "inception cohort," ie, a group of studies of a biomarker defined from the first evalu-

46

Publication bias and selective publication

THE NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

Erick H. Turner, M.D., Annette M. Matthews, M.D., Eftihia Linardatos, B.S., Robert A. Tell, L.C.S.W., and Robert Rosenthal, Ph.D.

ABSTRACT

BACKGROUND

Evidence-based medicine is valuable to the extent that the evidence base is complete and unbiased. Selective publication of clinical trials — and the outcomes within those trials — can lead to unrealistic estimates of drug effectiveness and alter the apparent risk–benefit ratio.

METHODS

We obtained reviews from the Food and Drug Administration (FDA) for studies of 12 antidepressant agents involving 12,564 patients. We conducted a systematic literature search to identify matching publications. For trials that were reported in the literature, we compared the published outcomes with the FDA outcomes. We also compared the effect size derived from the published reports with the effect size derived from the entire FDA data set.

RESULTS

Among 74 FDA-registered studies, 31%, accounting for 3449 study participants, were not published. Whether and how the studies were published were associated with the study outcome. A total of 37 studies viewed by the FDA as having positive results were published; 1 study viewed as positive was not published. Studies viewed by the FDA as having negative or questionable results were, with 3 exceptions, either not published (22 studies) or published in a way that, in our opinion, conveyed a positive outcome (11 studies). According to the published literature, it appeared that 94% of the trials conducted were positive. By contrast, the FDA analysis showed that 51% were positive. Separate meta-analyses of the FDA and journal data sets showed that the increase in effect size ranged from 11 to 69% for individual drugs and was 32% overall.

PLoS MEDICINE

Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration

ABSTRACT

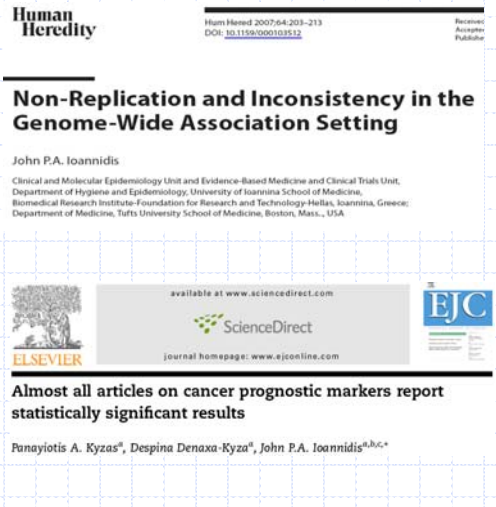
Background

Methods and Findings

Conclusions

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"47

Non-replicated studies and publication bias – especially in genetic and biomarker studies



Human Heredity

Hum Hered 2007;64:203–213
DOI: 10.1159/000103512

Non-Replication and Inconsistency in the Genome-Wide Association Setting

John P.A. Ioannidis

Clinical and Molecular Epidemiology Unit and Evidence-Based Medicine and Clinical Trials Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Biomedical Research Institute-Foundation for Research and Technology-Hellas, Ioannina, Greece; Department of Medicine, Tufts University School of Medicine, Boston, Mass., USA

ELSEVIER

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

EJC

Almost all articles on cancer prognostic markers report statistically significant results

Panayiotis A. Kyzas[a], Despina Denaxa-Kyza[a], John P.A. Ioannidis[a,b,c,*]

48

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2

PLoS Med 2005

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

TABLE 3. Characteristics of study quality

Characteristic	No. (%) of studies
Study design	
Cross-sectional	39 (15)
Case-control	208 (82)
Nested within observational study	7 (3)
Recruitment of participants	
Consecutive or random	20 (8)
Convenience or not reported	234 (92)
Selection criteria clearly described	141 (56)
Complete verification by use of the reference standard	107 (42)
Execution of test described in sufficient detail	253 (100)[a]
Index test results blinded to reference standard?	
Yes	65 (26)
No	1 (0)
Not reported	188 (74)

[a] The description of the test execution was deemed insufficient in one study.

Steingart KR et al.

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

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

51

Spectrum bias (a form of selection bias)

- Population used for evaluating the test:
  - 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

Clinical Chemistry 51:8
1335–1341 (2005)

Minireview

Case–Control and Two-Gate Designs in Diagnostic Accuracy Studies

Anne W.S. Rutjes,[1*] Johannes B. Reitsma,[1] Jan P. Vandenbroucke,[2] Afina S. Glas,[3] and Patrick M.M. Bossuyt[1]

52

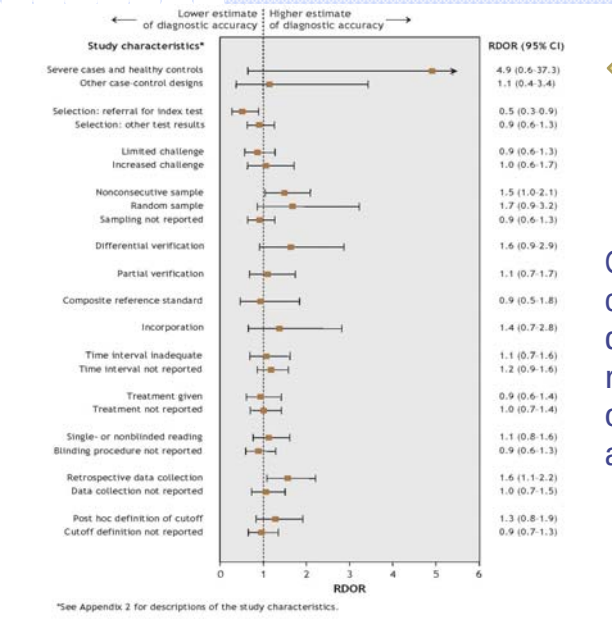


Fig. 2: Effects of study design characteristics on estimates of diagnostic accuracy. RDOR = relative diagnostic odds ratio (adjusted RDORs were estimated in a multivariable random-effects meta-epidemiologic regression model).

Case-control design results in optimistic accuracy

53

Ruties et al. CMAJ 2006

We find this in TB as well: Example: PCR tests for TB meningitis

Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis

Madhukar Pai, Laura L Flores, Nitika Pai, Alan Hubbard, Lee W Riley, and John M Colford Jr

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

Subgroup	Number of studies	Summary diagnostic odds ratio* (95% CI)	Test for heterogeneity† p value
Study design			
Case-control	19	86·5 (39·3, 190·2)	0·03
Cross-sectional	16	43·3 (22·5, 83·3)	0·94
Blinded interpretation of test and/or reference standard results			
Yes	21	46·9 (24·9, 88·6)	0·16
No	14	82·3 (39·8, 170·2)	0·70
Consecutive or random sampling of participants			
Yes	18	63·3 (32·8, 122·4)	0·20
No	17	46·8 (23·6, 92·8)	0·42
Prospective data collection			
Yes	18	59·9 (28·1, 127·6)	0·12
No	17	55·2 (29·9, 101·6)	0·59

*Random effects model. †χ^2 test for heterogeneity. CI=confidence interval.

54

Pai et al. Lancet Infect Dis 2003

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

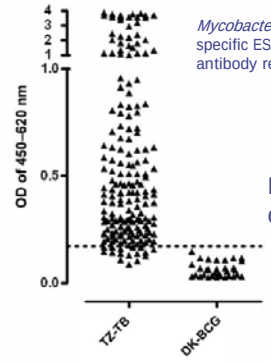
55

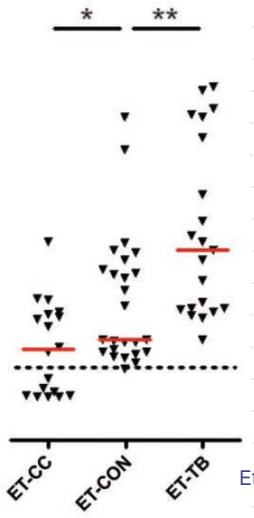
Lack of discrimination in TB endemic settings: example

Mycobacterium tuberculosis–specific ESAT-6 and CFP-10 antibody responses

Tanzanian cases

Danish controls





Ethiopia

Figure 3. Dot-plot showing the optical density (OD) values obtained from 184 patients with active tuberculosis disease who resided in northern Tanzania (TZ-TB) and 32 healthy, bacille Calmette-Guérin–vaccinated, Danish resident volunteer donors with no known risk factors for tuberculosis (DK-BCG). The dotted line indicates the cutoff value, calculated as the mean OD + 3 SDs for the 32 healthy Danish resident volunteers.

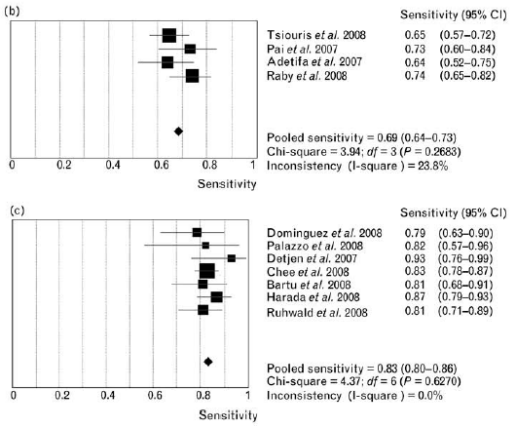
56

Hoff et al. Clinical Infectious Diseases 2007; 45:575–82

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[a,b,c], Richard van Zyl Smit[a], Motasim Badri[a] and Madhukar Pai[d]



High incidence countries

Low incidence countries

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

Evaluation of the MycoDot™ test in patients with suspected tuberculosis in a field setting in Tanzania

G. R. Somi,* R. J. O'Brien,† G. S. Mfinanga,* Y. A. Ipuge‡

*National Institute for Medical Research, Dar Es Salaam, Tanzania, †WHO Global Tuberculosis Programme, Geneva, Switzerland, ‡National Tuberculosis and Leprosy Programme, Dar Es Salaam, Tanzania

SUMMARY

SETTING: Rapid, simple and inexpensive methods are needed to improve the diagnosis of tuberculosis in low-income countries. The MycoDot™ test has these characteristics.

OBJECTIVE: To assess the utility of the MycoDot™ test in screening patients with suspected tuberculosis.

DESIGN: Ambulatory patients presenting with symptoms of pulmonary tuberculosis were evaluated by physical examination and sputum acid-fast bacilli (AFB) microscopy. Separately, the MycoDot™ test was performed on whole blood. Patients with AFB-negative smears were treated with a 10-day course of erythromycin. Those remaining symptomatic had a chest radiograph. All sputum specimens were cultured for mycobacteria. Patients with culture-negative tuberculosis and those without a tuberculosis diagnosis were reassessed at 2 months.

RESULTS: Among the 241 patients who were evaluated, the MycoDot™ test was positive in 26% of patients with AFB-positive/culture-positive tuberculosis, 7% with AFB-negative/culture-positive tuberculosis, 7% with culture-negative tuberculosis, 19% treated for tuberculosis who did not meet study case definitions, and 16% without tuberculosis. Twenty four patients did not complete the assessment. Test sensitivity was 16%, specificity 84% and positive predictive value 45%. Sensitivity was highest (41%) in AFB-positive/HIV-negative patients and lowest (3%) in AFB-negative/HIV-positive patients.

CONCLUSION: The MycoDot™ test is not useful for the diagnosis of tuberculosis in sub-Saharan African countries, especially where HIV infection is prevalent.

KEY WORDS: tuberculosis; diagnosis; HIV; serology

Sens in HIV+ = 26%

Evaluation of a commercial immunodiagnostic kit incorporating lipoarabinomannan in the serodiagnosis of pulmonary tuberculosis in Ghana

S. D. Lawn†, E. H. Frimpong‡ and E. Nyarko§

† Department of Medicine, School of Medical Sciences, University of Science and Technology, Kumasi, Ghana
‡ Department of Microbiology, School of Medical Sciences, University of Science and Technology, Kumasi, Ghana
§ National Tuberculosis Control Programme, Ministry of Health, Accra, Ghana

Summary

We evaluated 'Mycodot', a commercially marketed immunodiagnostic test for tuberculosis which detects antibodies to lipoarabinomannan antigen. Serum was tested from 52 patients with newly diagnosed smear-positive pulmonary tuberculosis, of whom 20 were HIV-positive and 32 HIV-negative. Control sera were taken from 40 patients of whom 20 had acute non-tuberculous lobar pneumonia and 20 patients had no respiratory disease. The test was found to have a very high specificity of 97.5% (95%CI 92.5–100%). However, the sensitivity in HIV-negative patients was 56% (95%CI 39–73%), and was substantially lower at 25% (95%CI 6–44%) in HIV-positive patients. In conclusion: 'Mycodot' was found to be a highly specific and easily performed assay. However, the poor sensitivity, especially in HIV- infected patients, renders it unlikely to be useful either as a primary or adjunctive diagnostic test for tuberculosis, particularly in countries with a high prevalence of HIV. A larger trial of this assay in Ghana was not deemed necessary.

Sens in HIV+ = 25%

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

59

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

60

Example: exclusion of indeterminates can inflate accuracy estimates

OPEN ACCESS Freely available online

PLoS one

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

Basirudeen Syed Ahamed Kabeer[1], Rajasekaran Sikhamani[5], Sowmya Swaminathan[2], Venkatesan Perumal[3], Paulkumaran Paramasivam[4], Alamelu Raja[1]*

1 Department of Immunology, Tuberculosis Research Centre (ICMR), Tamil Nadu, India, 2 Division of HIV/AIDS, Tuberculosis Research Centre (ICMR), Tamil Nadu, India, 3 Department of Statistics, Tuberculosis Research Centre (ICMR), Tamil Nadu, India, 4 Department of Clinic, Tuberculosis Research Centre (ICMR), Tamil Nadu, India, 5 Government Hospital of Thoracic Medicine, Tambaram Sanatorium, Tambaram, Tamil Nadu, India

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).

Methodology/Principal Findings: A total of 105 HIV-TB patients who were naive for anti tuberculosis and anti retroviral therapy were included for this study out of which 53 (50%) were culture positive. Of 105 tested, QuantiFERON-TB Gold in-tube (QFT-G) was positive in 65% (95% CI: 56% to 74%), negative in 18% (95% CI: 11% to 25%) and indeterminate in 17% (95% CI: 10% to 24%) of patients. The sensitivity of QFT-G remained similar in pulmonary TB and extra-pulmonary TB patients. The QFT-G positivity was not affected by low CD4 count, but it often gave indeterminate results especially in individuals with CD4 count <200 cells/μl. All of the QFT-G indeterminate patients whose sputum culture were positive, showed ≤0.25 IU/ml of IFN-γ response to phytohemagglutinin (PHA). TST was performed in all the 105 patients and yielded the sensitivity of 31% (95% CI: 40% to 22%). All the TST positives were QFT-G positives. The sensitivity of TST was decreased, when CD4 cell counts declined.

Conclusions/Significance: Our study shows neither QFT-G alone or in combination with TST can be used to exclude the suspicion of active TB disease. However, unlike TST, QFT-G yielded fewer false negative results even in individuals with low CD4 count. The low PHA cut-off point for indeterminate results suggested in this study (≤0.25 IU/ml) may improve the proportion of valid QFT-G results.

Citation: Syed Ahamed Kabeer B, Sikhamani R, Swaminathan S, Perumal V, Paramasivam P, et al. (2009) Role of Interferon Gamma Release Assay in Active TB Diagnosis among HIV Infected Individuals. PLoS ONE 4(5): e5718. doi:10.1371/journal.pone.0005718

- If indeterminates are included:
  - Sens = 66%
- If indeterminates are excluded:
  - Sens = 85%

61

J Clin Epidemiol Vol. 52, No. 12, pp. 1231–1237, 1999
Published by Elsevier Science Inc.

ELSEVIER

0895-4356/99/$–see front matter
PII S0895-4356(99)00101-8

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

*Alula Hadgu**

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

ABSTRACT. Discrepant 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 discrepant analysis–based estimates of specificity are typically less biased than those based on culture and that the discrepant analysis–based specificity shows little appreciable bias. In this article, I show that those conclusions are incorrect. Using a typical example from the published literature, I show that the discrepant 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 discrepant analysis is profoundly flawed and unscientific. It violates a fundamental principle of diagnostic testing—the principle that the new test should not be used to determine the true disease status. Thus, the major problem with discrepant analysis is not only that it is biased but that it is unscientific. Therefore, discrepant analysis should not be adopted for the evaluation of any diagnostic or screening test. J CLIN EPIDEMIOL 52;12:1231–1237, 1999. Published by Elsevier Science Inc.

KEY WORDS. Discrepant analysis, sensitivity, specificity, DNA-amplification tests, *Chlamydia trachomatis*

62

Industry involvement in drug trials and its impact on study outcomes and conclusions

Scope and Impact of Financial Conflicts of Interest in Biomedical Research
A Systematic Review

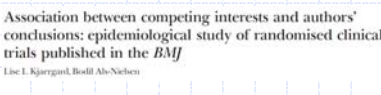
JAMA 2003

Pharmaceutical industry sponsorship and research outcome and quality: systematic review

BMJ 2003

Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials

CMAJ 2004

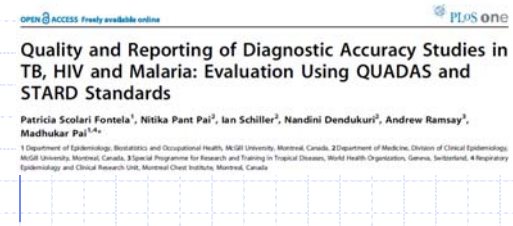


Association between competing interests and authors' conclusions: epidemiological study of randomised clinical trials published in the *BMJ*

BMJ 2002

63

Industry involvement in diagnostic studies?



OPEN ACCESS Freely available online PLoS one

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

Patricia Scolari Fontela[1], Nitika Pant Pai[2], Ian Schiller[2], Nandini Dendukuri[2], Andrew Ramsay[3], Madhukar Pai[1,4]*

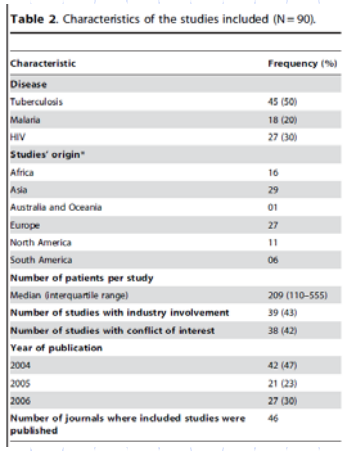


Table 2. Characteristics of the studies included (N = 90).

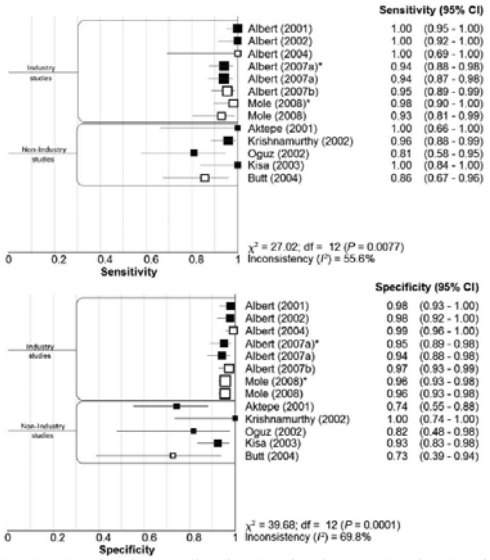
Characteristic	Frequency (%)
Disease	
Tuberculosis	45 (50)
Malaria	18 (20)
HIV	27 (30)
Studies' origin*	
Africa	16
Asia	29
Australia and Oceania	01
Europe	27
North America	11
South America	06
Number of patients per study	
Median (interquartile range)	209 (110–555)
Number of studies with industry involvement	39 (43)
Number of studies with conflict of interest	38 (42)
Year of publication	
2004	42 (47)
2005	21 (23)
2006	27 (30)
Number of journals where included studies were published	46

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

Fontela P et al. PLoS One 2009

64

FASTPlaque tests for drug-resistant TB



Minion J et al. ITJLD 2010 65

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

SEROCHECK-MTB	Rapid Test for Antibodies to *Mycobacterium tuberculosis* in serum/ plasma/whole blood
Application	• As a additional diagnostic tool in tuberculosis smear negative, culture positive suspects and tuberculosis smear negative, culture negative suspects • Extrapulmonary TB suspects • Pediatric cases • Diagnosis of suspect TB cases in HIV uninfected individuals
Principle	Self performing, rapid , semi-quantitative two-site sandwich immunoassay , lateral flow device
Sensitivity	100%
Specificity	100%

Sensitivity = 100%
Specificity = 100%

SD STANDARD DIAGNOSTICS, INC.
156-68 Hagal-dong, Giheung-gu, Yongin-si, Kyunggi-do, Korea
Tel : 82-31-899-9700 Fax : 82-31-899-9743 http://www.standardia.com

2) Comparison SD Rapid TB vs. a commercial anti-TB ELISA
The SD Rapid TB have tested with positive and negative clinical samples tested by a leading commercial ELISA test. The result shows that the SD Rapid TB is very accurate to other commercial ELISA test.

		A Commertial PHA		Total Results
		Positive	Negative	
A commercial anti-Mtuberculosis ELISA kit	Positive	112	2	114
	Negative	1	350	351
Total Results		113	352	465

In a comparison of the SD Rapid TB versus a leading commercial ELISA test, results gave sensitivity of 98.2% (112/114), a specificity of 99.7 % (350/351), and a total agreement of 99.35% (462/465).

Sensitivity = 98%
Specificity = 100%

TASHIMA Inc.

PERFOMANCE CHARACTERISTICS:

Sensitivity : Sera were collected from patients under anti TB treatment. Results of sputum examination were not available. Among 75 sera collected, samples were positive by the TB onsite Rapid screening Test Thus, the test sensitivity is 93%.

Specificity : In 53 sera derived from Northern America, all the samples were negative.

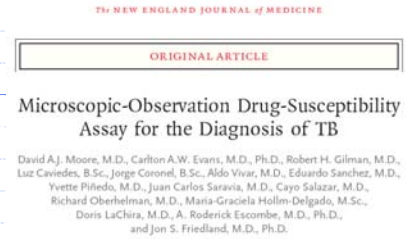
Sensitivity = 93%
Specificity = 100%

66

More examples…

Test	Package insert sens	Package insert spec	Meta-analysis sens	Meta-analysis spec
QFT-Gold	89%		79%	
FASTPlaque-Response	96 – 100%	99 – 100%	95%	95%
Anda-TB IgG	85 - 90%	85 - 100 %	60 - 75%	~90%
MycoDot	70%	95%	26% - 76%	84% - 97%
Clearview TB ELISA	81% (HIV+)	93 – 98%	56% (HIV+)	95%
GenoType MDTBDrplus	99%	99%	98%	99%
Gen-Probe MTD	97% (S+) 72 (S-)	100% (S+) 99% (S-)	97% (S+) 76% (S-)	96% (S+) 95% (S-)

67


The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

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

David A.J. Moore, M.D., Carlton A.W. Evans, M.D., Ph.D., Robert H. Gilman, M.D., Luz Caviedes, B.Sc., Jorge Coronel, B.Sc., Aldo Vivar, M.D., Eduardo Sanchez, M.D., Yvette Piñedo, M.D., Juan Carlos Saravia, M.D., Cayo Salazar, M.D., Richard Oberhelman, M.D., Maria-Graciela Hollm-Delgado, M.Sc., Doris LaChira, M.D., A. Roderick Escombe, M.D., Ph.D., and Jon S. Friedland, M.D., Ph.D.

MODS: developed in Peru – performs excellent

Sensitivity better than LJ (98 vs. 84%)

Fast turnaround time (1 week vs. 6 weeks+)

Implemented in India – performs poorly

Sensitivity 80%

Issues with contamination

Issues with reliability


INT J TUBERC LUNG DIS 14(4):482–488
©2010 The Union

Diagnostic accuracy of the microscopic observation drug susceptibility assay: a pilot study from India

J. S. Michael,* P. Daley,† S. Kalaiselvan,* A. Latha,† J. Vijayakumar,† D. Mathai,† K. R. John,‡ M. Pai§

68

Simple, phage-based (*FASTPlaque*) technology to determine rifampicin resistance of *Mycobacterium tuberculosis* directly from sputum

H. Albert,* A. Trollip,* T. Seaman,* R. J. Mole*

* Biotec Laboratories Ltd, c/o National Health Laboratory Service, Cape Town, Western Cape, South Africa; † Biotec Laboratories Ltd., Ipswich, Suffolk, United Kingdom

SUMMARY

SETTING: Cape Town, South Africa.
OBJECTIVE: To evaluate the performance of a simple, manual, phage-based test for determining rifampicin (RMP) resistance of *Mycobacterium tuberculosis* directly from smear-positive sputum specimens.
DESIGN: A comparative study of the performance of the *FASTPlaque* (phage amplification) technology to determine RMP resistance directly from smear-positive sputum compared with isolation and the conventional indirect Middlebrook 7H11 agar proportion method.
RESULTS: The *FASTPlaque* direct RMP test achieved sensitivity, specificity and overall accuracy of 100% (11/11), 100% (134/134) and 100% (145/145), respectively, compared with the conventional indirect susceptibility test method (resolved data). The *FASTPlaque* direct RMP test reported results within 2 days from receipt of the specimen, while the conventional method took between 27 and 103 days (mean ± SD 33.2 ± 7.2 days).
CONCLUSION: *FASTPlaque* technology applied directly to smear-positive sputum offers performance comparable to conventional methods, with results available in 2 days instead of weeks to months. The test may form a useful part of DOTS-Plus programmes to combat multidrug-resistant tuberculosis, improving patient prognosis and reducing ongoing transmission of disease. It does not require specialised equipment, making it appropriate for high-burden countries.
KEY WORDS: *FASTPlaque*; phage amplification; multidrug-resistant tuberculosis; rifampicin resistance; susceptibility testing

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)

Evaluation of *FASTPlaque*TB™ to diagnose smear-negative tuberculosis in a peripheral clinic in Kenya

M. Bonnet,* L. Gagnidze,* F. Varaine,† A. Ramsay,‡§ W. Githui,¶ P. J. Guerin*

* Epicentre, Paris, † Médecins Sans Frontières, Paris, France; ‡Liverpool School of Tropical Medicine, Liverpool, UK; §United Nations Children's Fund/United Nations Development Programme/World Bank/World Health Organization Special Programme for Research and Training for Tropical Diseases (TDR), Geneva, Switzerland; ¶Centre for Respiratory Diseases Research, Kenya Medical Research Institute, Nairobi, Kenya

SUMMARY

OBJECTIVE: To evaluate the performance and feasibility of *FASTPlaque*TB™ in smear-negative tuberculosis (TB) suspects in a peripheral clinic after laboratory upgrading.
DESIGN: Patients with cough >2 weeks, two sputum smear-negative results, no response to 1 week of amoxicillin and abnormal chest X-ray were defined as smear-negative suspects. One sputum sample was collected, decontaminated and divided into two: half was tested with *FASTPlaque*TB in the clinic laboratory and the other half was cultured on Löwenstein-Jensen medium in the Kenyan Medical Research Institute. Test sensitivity and specificity were evaluated in all patients and in human immunodeficiency virus (HIV) infected patients. Feasibility was assessed by the contamination rate and the resources required to upgrade the laboratory.
RESULTS: Of 208 patients included in the study, 56.2% were HIV-infected. Of 203 *FASTPlaque*TB tests, 95 (46.8%) were contaminated, which interfered with result interpretation and led to the interruption of the study. Sensitivity and specificity were respectively 31.2% (95%CI 12.1–58.5) and 94.9% (95%CI 86.8–98.4) in all patients and 33.3% (95%CI 9.9–65.1) and 93.9% (95%CI 83.1–98.7) in HIV-infected patients. Upgrading the laboratory cost €20 000.
CONCLUSION: *FASTPlaque*TB did not perform satisfactorily in this setting. If contamination can be reduced, in addition to laboratory upgrading, its introduction in peripheral clinics would require further assessment in smear-negative and HIV co-infected patients and test adaptation for friendlier use.
KEY WORDS: tuberculosis; phage-based test; smear microscopy; diagnosis; developing countries

69

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

INT J TUBERC LUNG DIS 2(7):541-546
© 1998 IUATLD

MPB64 mycobacterial antigen: a new skin-test reagent through patch method for rapid diagnosis of active tuberculosis

R. M. Nakamura,* M. A. Velmonte,† K. Kawajiri,* C. F. Ang,† R. A. Frias,† M. T. Mendoza,† J. C. Montoya,† I. Honda,* S. Haga,‡ I. Toida*

* Japan BCG Laboratory, Kiyose-shi, Tokyo, Japan, †Infectious Disease Section, Philippine General Hospital, Manila, Philippines, ‡National Institute of Infectious Diseases, Toyama, Shinjuku-ku, Tokyo, Japan

SUMMARY

SETTING: A collaborative study between the Japan BCG Laboratory, Tokyo, Japan, and the Infectious Disease Section, Philippine General Hospital, Manila, the Philippines. Tuberculosis patients from four clinics in the vicinity of Manila, Our Lady of Grace Parish, Sto. Niño de Tondo Parish, the Canossa Health and Social Center, and the Health Care Development Center, were examined.
OBJECTIVE: To develop a new, simple and rapid diagnostic method for active tuberculosis. Subjects were tested for skin reaction to a special antigen, MPB64, by the patch test method instead of intradermal injection of purified protein derivative (PPD).
DESIGN: Fifty-three active tuberculosis patients and 43 healthy PPD-positive controls were tested to determine whether or not the reaction to MPB64 was positive only in active tuberculosis patients.
RESULTS: Fifty-two of the 53 active tuberculosis patients showed a positive reaction to MPB64, while none of the 43 PPD-positive controls did. The specificity of MPB64 to active tuberculosis was 100%, and the sensitivity was 98.1%. The efficacy of the test was 98.9%.
CONCLUSION: The patch test with MPB64 is a promising method for the diagnosis of active tuberculosis, distinguishing tuberculous patients from those who are infected but have not developed the disease, and also from BCG-vaccinated individuals. This new skin test is a subject for further evaluation and it is important to compare the results with PPD Mantoux.
KEY WORDS: MPB64; patch skin test; rapid diagnosis; active TB

Early data in 1998:

Sensitivity: 98%
Specificity: 100%

In 2012, still not commercially available

70

Pai M et al. Exp Rev Mol Diagn 2006;6(3):423-432; Image courtesy Sequella Inc.