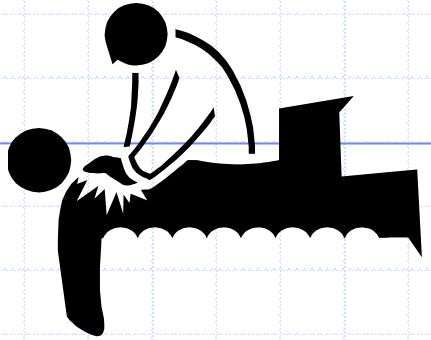
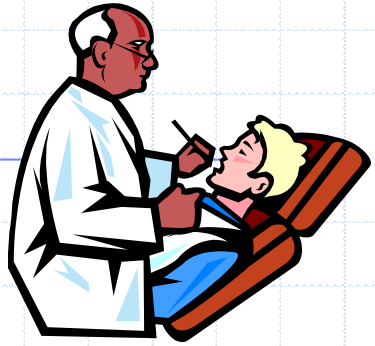
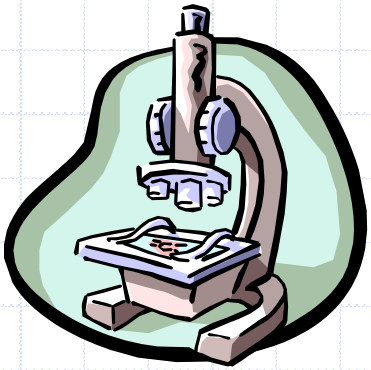
Diagnostic research designs: an introductory overview







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

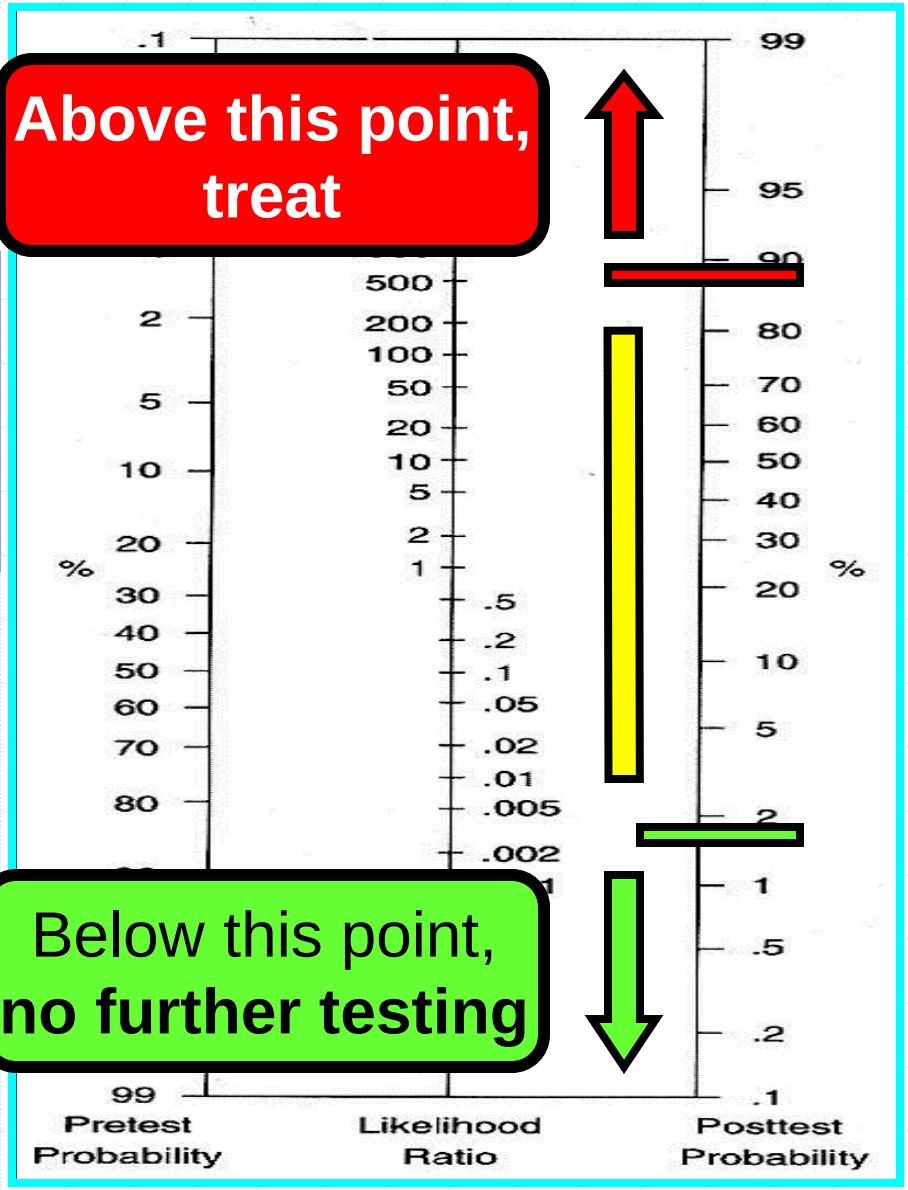
Email: madhukar.pai@mcgill.ca

Approaches to Diagnosis

Pattern recognition	Probabilistic diagnostic reasoning
See it and recognize disorder	Clinical assessment generates pretest probability
↓	↓
Compare posttest probability with thresholds	New information generates posttest probability
(usually pattern recognition implies probability near 100% and so above threshold)	(may be interative)
	↓
	Compare posttest probability with thresholds

Source: Guyatt G, Rennie D, Meade MO, Cook DJ: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*, 2nd Edition: http://www.jamaevidence.com

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



Disease ruled IN

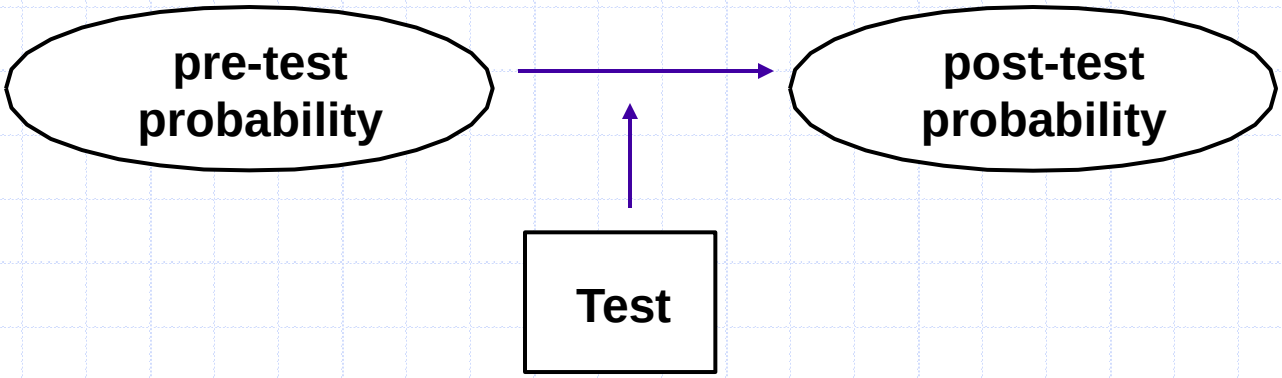
Disease not ruled in or out

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

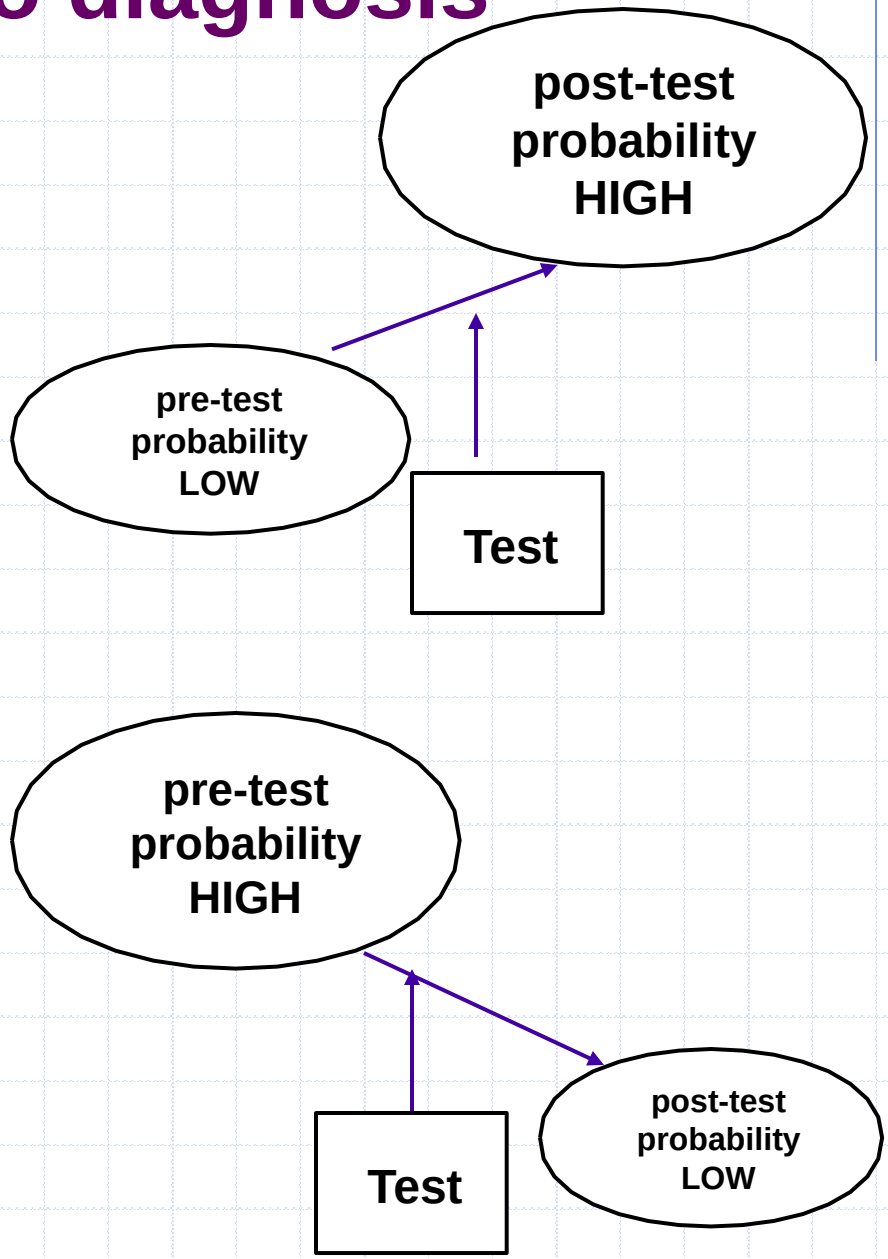
What you thought before + New information = What you think now



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

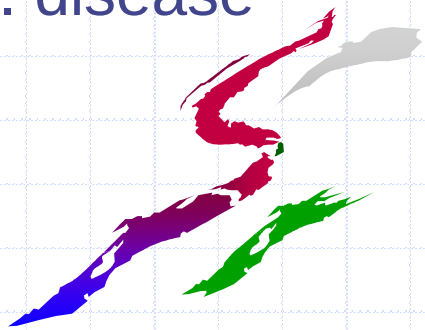
Opinion

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

FRAMINGHAM HEART STUDY

A Project of the National Heart, Lung and Blood Institute and Boston University

About FHS | Participants | FHS Investigators | Risk Score Profiles | FHS Bibliography | For Researchers

- Atrial Fibrillation (10-year risk)
- Cardiovascular Disease (30-year risk)
- Congestive Heart Failure
- Coronary Heart Disease (10-year risk)
- Coronary Heart Disease (2-year risk)
- Diabetes Risk Score
- General Cardiovascular Disease (10-year risk)
- Hard Coronary Heart Disease (10-year risk)
- Hypertension Risk Score
- Intermittent Claudication
- Recurring Coronary Heart Disease
- Stroke
- Stroke After Atrial Fibrillation
- Stroke or Death

Cardiovascular Disease (30-year risk)

(based on Pencina, D'Agostino, Larson, Massaro, Vasan. 'Predicting the 30-Year Risk of Cardiovascular Disease: The Framingham Heart Study', Circulation 2009)

Outcome

"Hard" CVD (coronary death, myocardial infarction, stroke), "general" CVD (coronary death, myocardial infarction, coronary insufficiency, angina, ischemic stroke, hemorrhagic stroke, transient ischemic attack, peripheral artery disease, heart failure)

Duration of follow-up

Maximum of 35 years, 30-year risk prediction

Population of interest

Individuals 20 to 59 years and free of CVD and cancer at baseline examination

Predictors

- Male Sex
- Age
- Systolic Blood Pressure (SBP)
- Use of Antihypertensive treatment (yes/ no)
- Smoking
- Diabetes mellitus
- Total cholesterol
- HDL cholesterol
- BMI replacing lipids in a simpler model

Risk Score Calculator

We acknowledge Mr. Aaron Vaneps and the Mayo Clinic Cardiovascular Health Clinic who provided the interactive risk calculator.

30 Year Risk Factors

Sex:	(•) Male ( ) Female
Systolic BP:	
Age:	
Diabetes:	☐
Smoker:	☐
Treated Hypertension:	☐
Total Cholesterol:	
HDL Cholesterol:	
BMI:	

National Cancer Institute

U.S. National Institutes of Health | www.cancer.gov

Breast Cancer Risk Assessment Tool

An interactive tool to help estimate a woman's risk of developing breast cancer

Last modified date: 05/16/2011

- Risk Calculator
- About the Tool
- Breast Cancer Risk
- Mobile Access
- Download Source Code

Page Options

Print Page

Email Page

Quick Links

- Breast Cancer Home Page
- Breast Cancer: Prevention, Genetics, Causes
- Initial Results of STAR Released
- Current Clinical Trials: Breast Cancer *In Situ*: Treatment
- Current Clinical Trials: Breast Cancer Prevention
- Current Clinical Trials: Breast Cancer Screening
- Estimating Breast Cancer: Q&A
- Understanding Cancer Risk
- National Cancer Institute

Need Help?
Contact us by phone, Web, and e-mail
1-800-4-CANCER

The Breast Cancer Risk Assessment Tool is an interactive tool designed by scientists at the National Cancer Institute (NCI) and the National Surgical Adjuvant Breast and Bowel Project (NSABP) to estimate a woman's risk of developing invasive breast cancer. The tool has been updated for African American women based on the Contraceptive and Reproductive Experiences (CARE) Study, and for Asian and Pacific Islander women in the United States based on the Asian American Breast Cancer Study (AABCS). See About the Tool for more information.

Before using the tool, please note the following:

- The Breast Cancer Risk Assessment Tool was designed for use by health professionals. If you are not a health professional, you are encouraged to discuss the results and your personal risk of breast cancer with your doctor.
- The tool should not be used to calculate breast cancer risk for women who have already had a diagnosis of breast cancer, lobular carcinoma in situ (LCIS), or ductal carcinoma in situ (DCIS).
- The BCRA risk calculator may be updated periodically as new data or research becomes available.
- Although the tool has been used with success in clinics for women with strong family histories of breast cancer, more specific methods of estimating risk are appropriate for women known to have breast cancer-producing mutations in the BRCA1 or BRCA2 genes.
- Other factors may also affect risk and are not accounted for by the tool. These factors include previous radiation therapy to the chest for the treatment of Hodgkin lymphoma or women who have recently immigrated to the United States from certain regions of Asia where breast cancer risk is low. Further, the tool may not be appropriate for women living outside the United States. The tool's risk calculations assume that a woman is screened for breast cancer as in the general U.S. population. A woman who does not have mammograms will have somewhat lower chances of a diagnosis of breast cancer.
- For information to help your patients understand cancer risk visit http://understandingrisk.cancer.gov. This interactive Web site will help your patients make informed decisions about how to lower their risk.

Risk Calculator

(Click a question number for a brief explanation, or read all explanations.)

1. Does the woman have a medical history of any breast cancer or of ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS)? — Select
2. What is the woman's age? *This tool only calculates risk for women 35 years of age or older.* — Select
3. What was the woman's age at the time of her first menstrual — Select

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.

Ann Intern Med. 2008;149:200-203. www.annals.org
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

Several other such classification schemes...

Freedman et al. 1987
I. Feasibility, promise
II. Studies of diagnostic accuracy
III. Studies of clinical value
IV. Studies for monitoring routine use

Thornbury and Fryback, 1992
I. Technical efficacy
II. Diagnostic accuracy efficacy
III. Diagnostic thinking efficacy
IV. Therapeutic efficacy
V. Patient outcome efficacy
VI. Societal efficacy

Pepe, 2005
I. Preclinical exploratory
II. Clinical assay and validation
III. Retrospective longitudinal
IV. Prospective screening
V. Disease control

Phased evaluation of medical tests

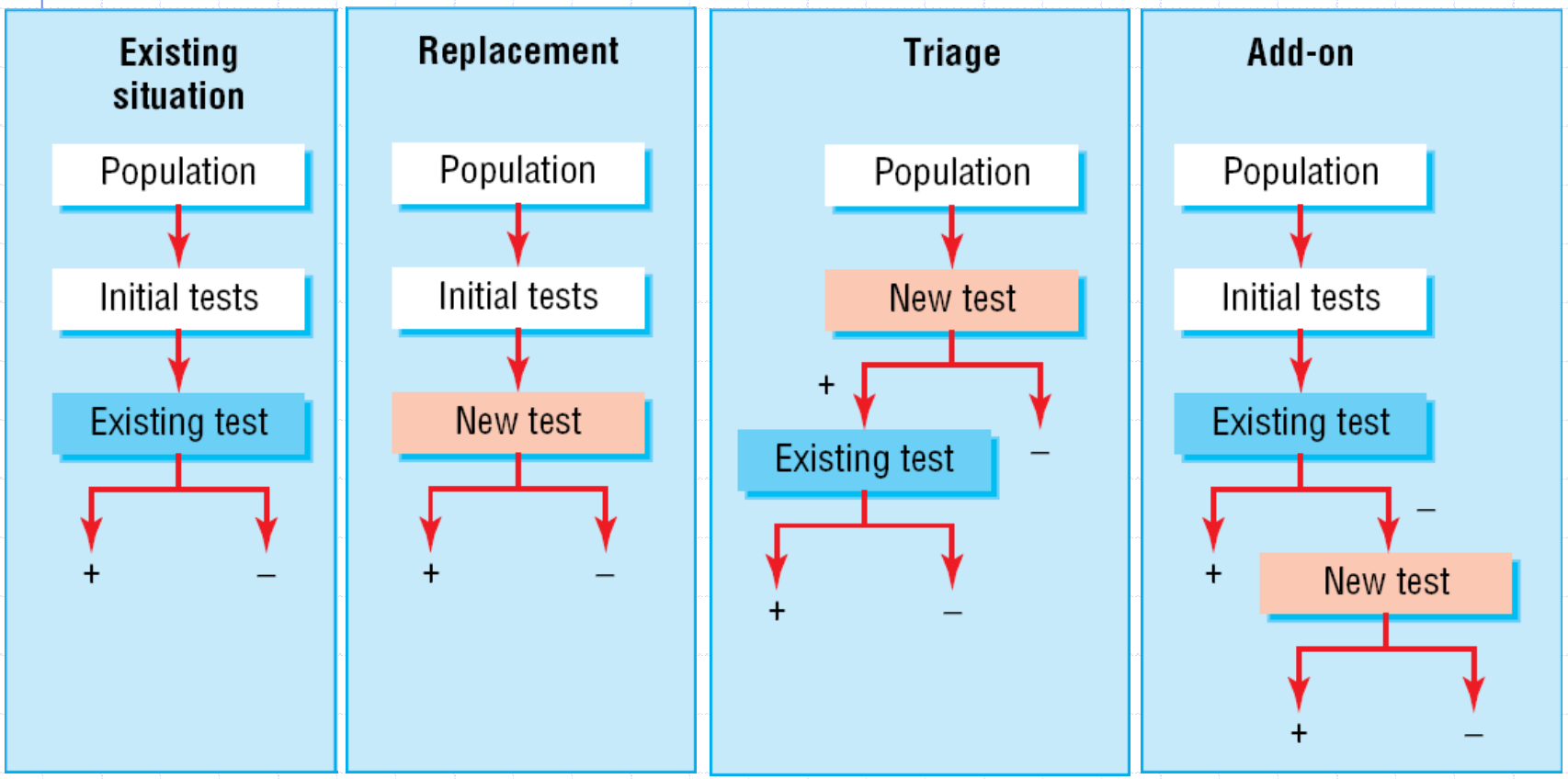
Levels/Phases
Technical efficacy
Intended use
Diagnostic accuracy
Usual range
Subgroups
Clinical population
Diagnostic thinking efficacy
Therapeutic efficacy
Patient outcome efficacy
Societal efficacy

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

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

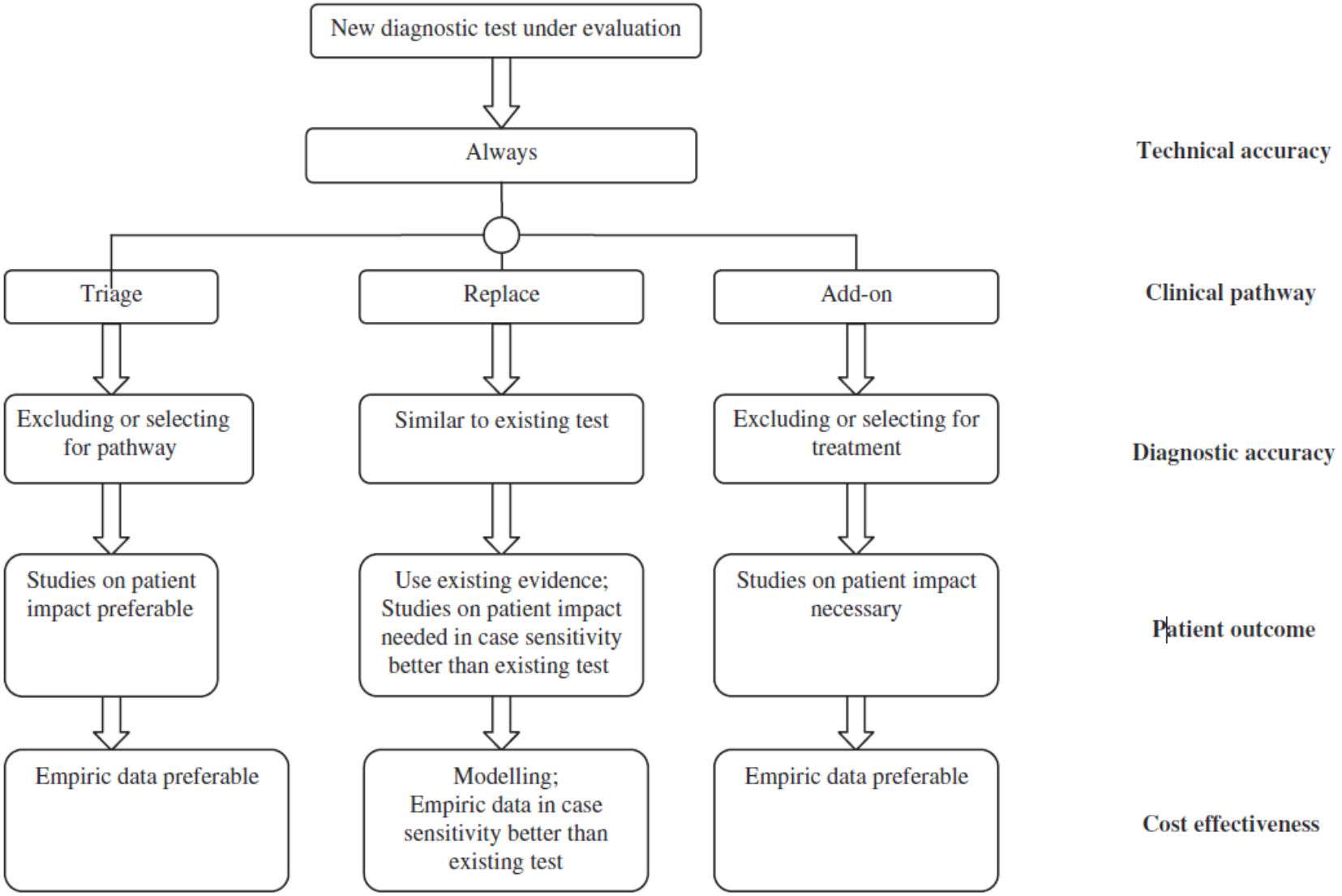
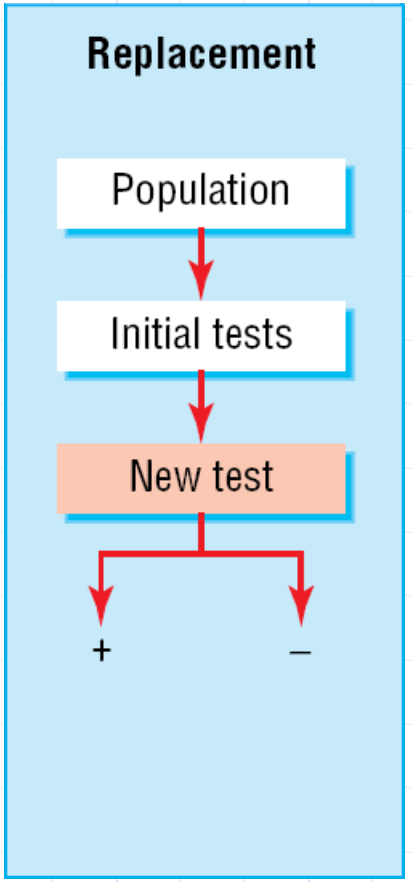


Fig. 1. Stepwise evaluation.

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...



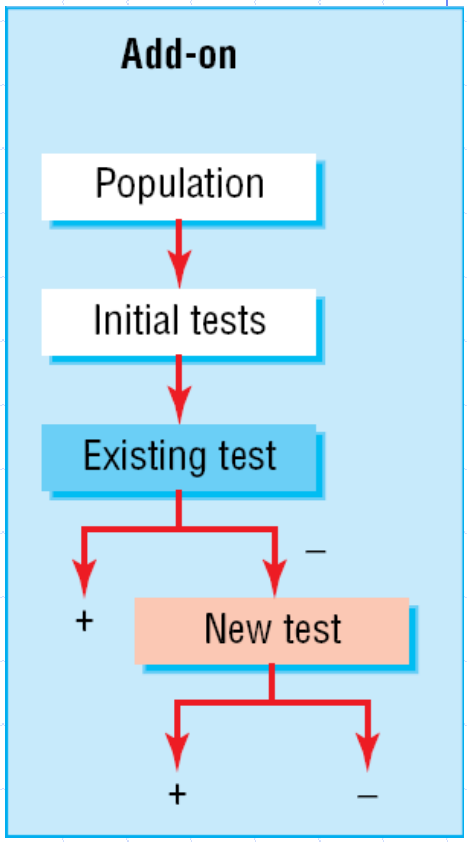
Add on

- Potential change in consequences, also extra numbers

  (either extra positives or extra negatives)

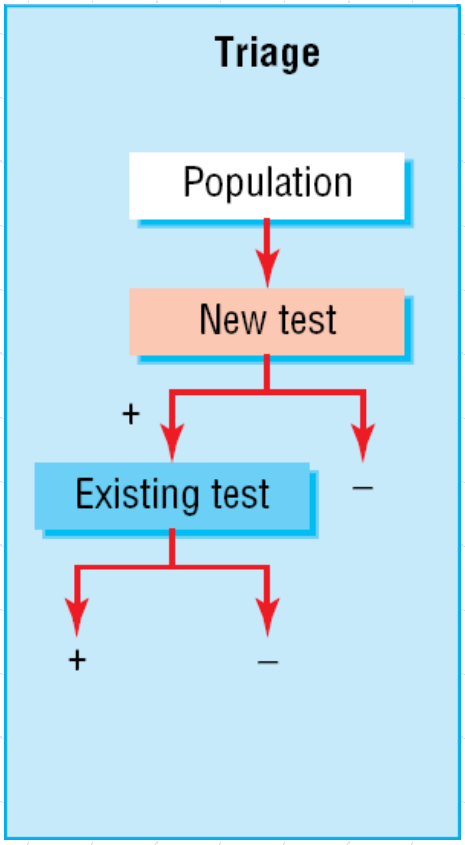
- Extra testing: extra time, burden

- Other info needed: costs, safety, burden, indeterminate results...
- Effect of change in consequences (patient impact)

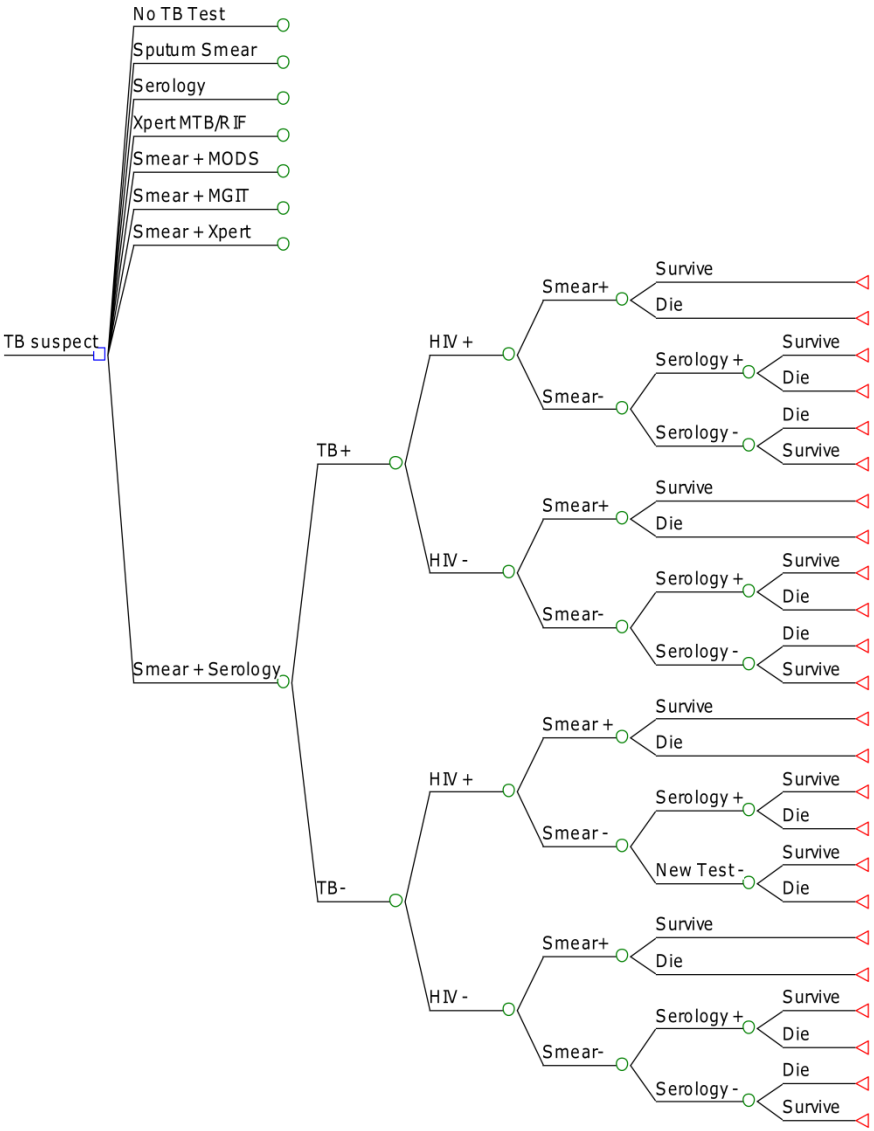


Triage

- 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?

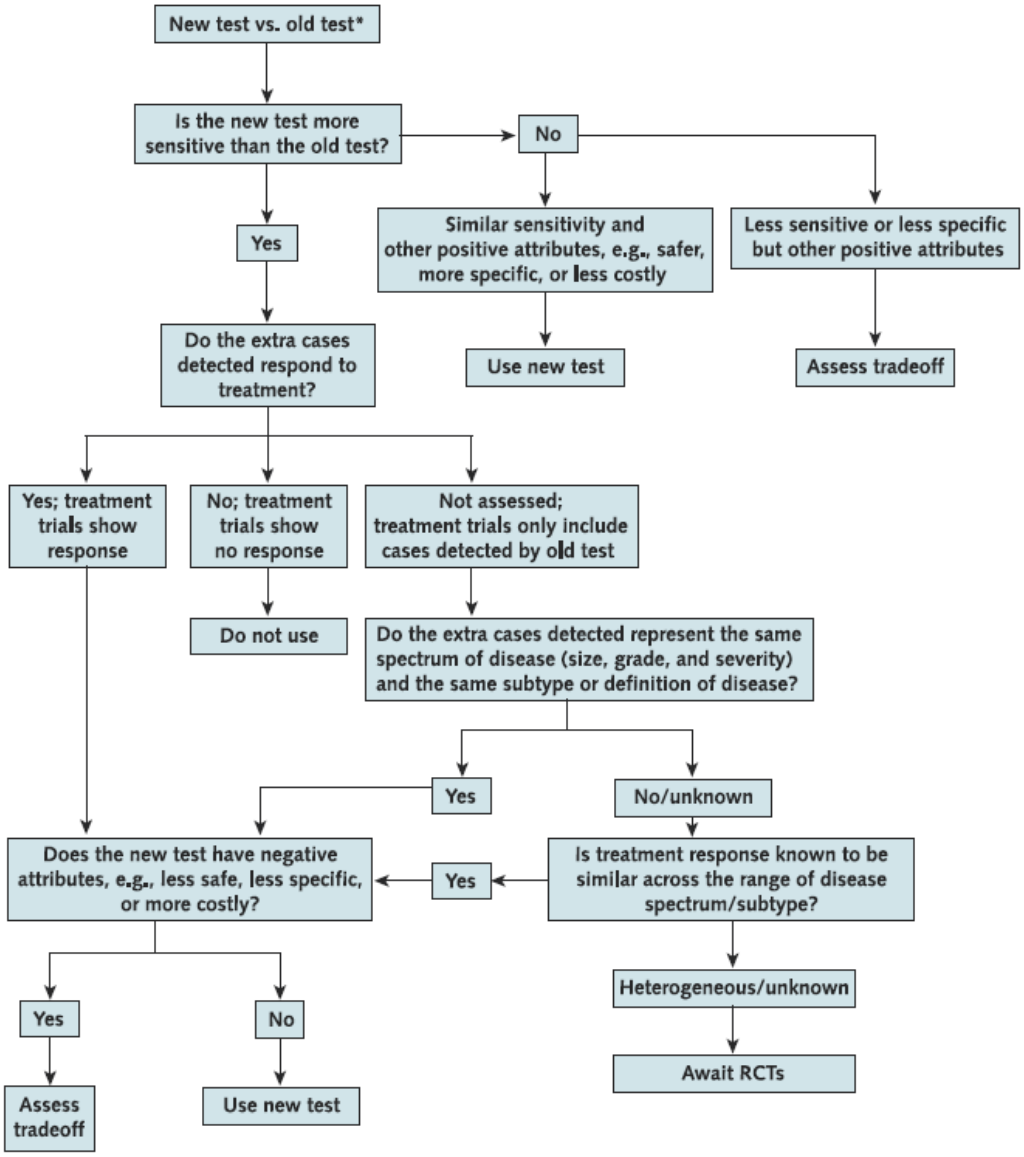
Sarah J. Lord, MBBS, MS; Les Irwig, MBBCh, PhD; and R. John Simes, MBBS, MS, MD

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, PhD§; 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 FluOIA 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.[2]

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 86 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

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

Methods

- Map the landscape of current TB diagnostic research
  - Bibliometric analysis of citations
  - PubMed and EMBASE were searched by a librarian for all original TB citations in a two year period - 2007-2008
    - For PubMed, the search strategy was: ("Mycobacterium tuberculosis"[Majr] OR "Tuberculosis"[Majr] OR "Tuberculosis/diagnosis"[Mesh] OR tuberculosis Field: Title) Limits: Publication Date from 2007/01/01 to 2008/12/31 NOT Field: Title, Editorial, Letter, Meta-Analysis, Practice Guideline, Review, Addresses, Bibliography, Biography, Comment, Dictionary, Directory, Interview, Newspaper Article.
    - For EMBASE, the search strategy was: exp *Mycobacterium Tuberculosis/ or exp *Tuberculosis or exp Tuberculosis/di [Diagnosis] or tuberculosis.m_titl. limit to yr="2007 - 2008“ not (book or book series or editorial or letter or "review")

Methods

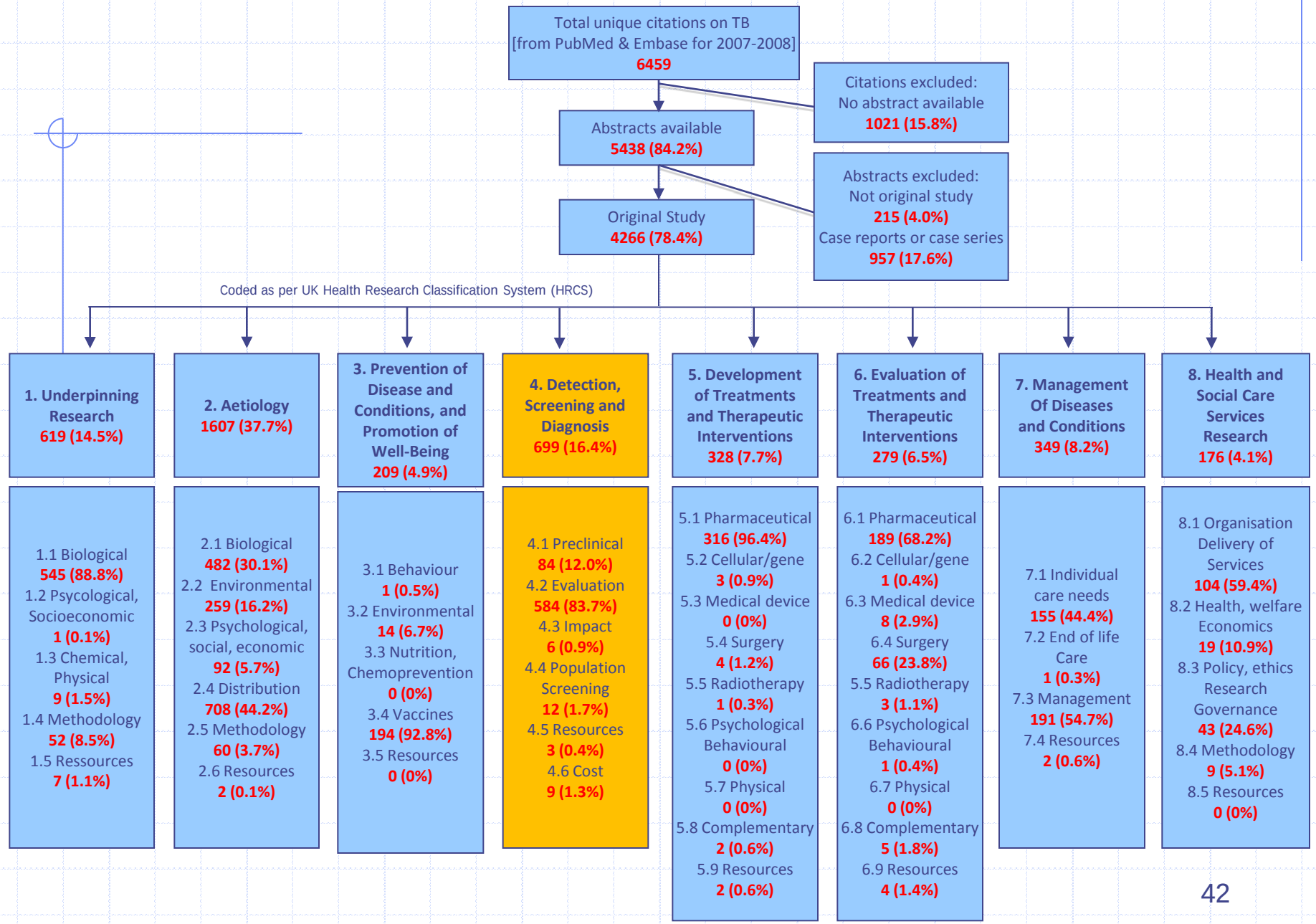
- Map the landscape of current TB diagnostic research
  - All the citations (titles and abstracts) were read and coded by a trained researcher after pilot testing and standardization
  - A second reviewer coded a subset of the citations
  - UK Clinical Research Collaboration's Health Research Classification System (HRCS) was used to retrieve details on the type of research of each study.
  - Additional information was collected for the diagnosis studies on: study design and type of outcome reported, purpose of the test, technology platform, study participants, study population, reporting of HIV status, use of commercial vs. in-house test, country where study was done, etc.

Methods

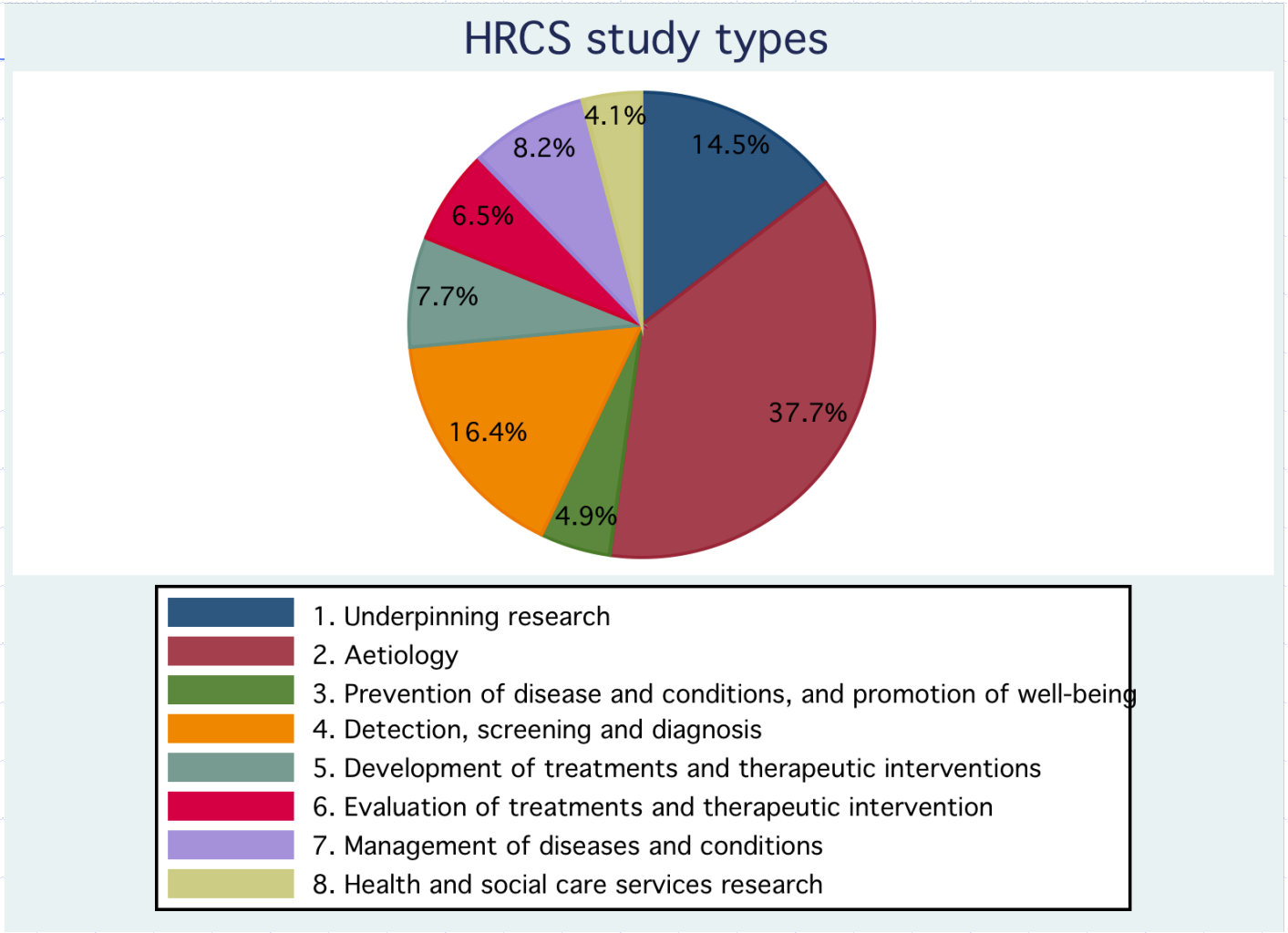
- Assess the quality of TB diagnostic accuracy studies

  - We used QUADAS and STARD checklists to assess the methodological and reporting quality of TB diagnostic studies published in a two year period

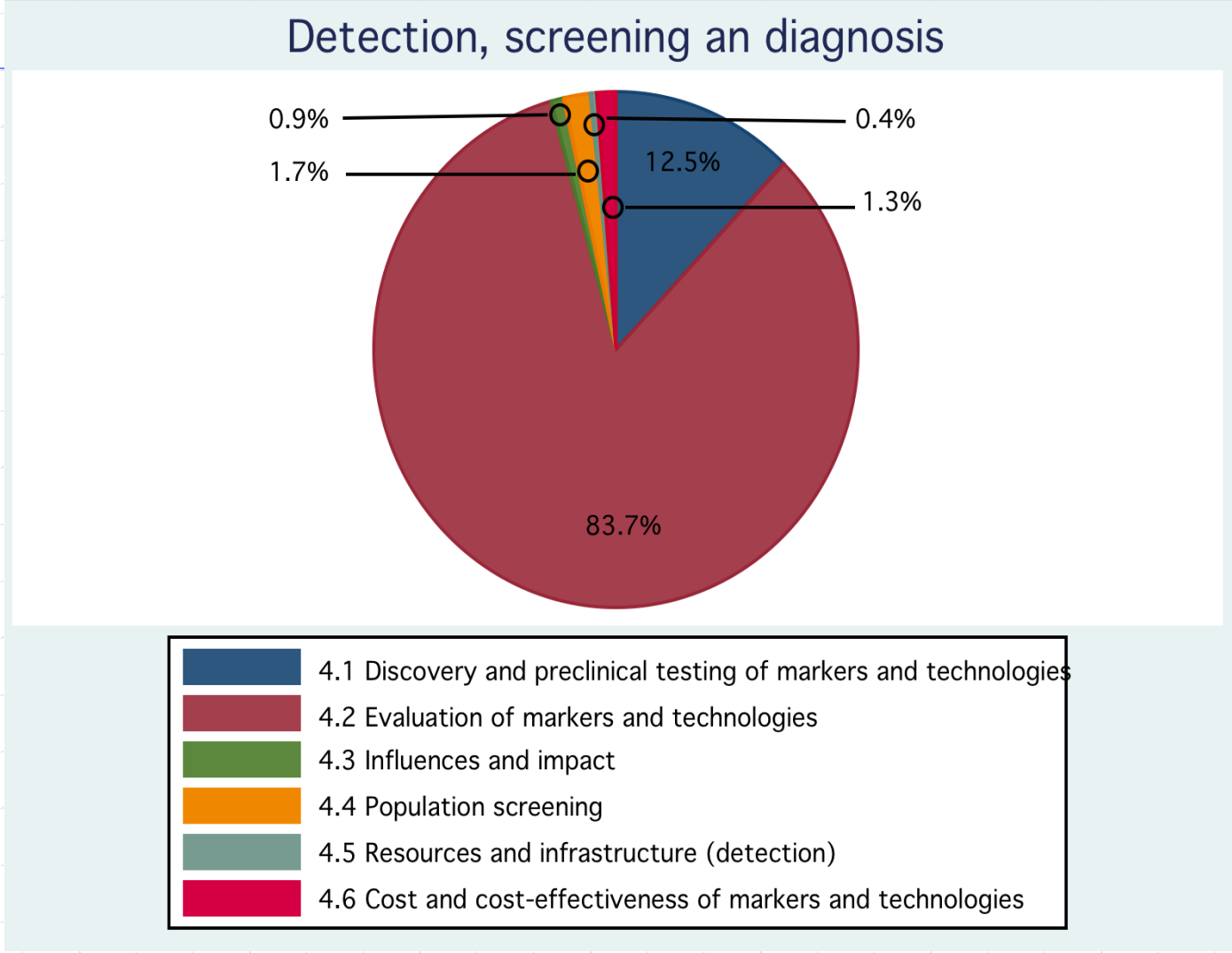
  - We also used several diagnostic meta-analyses to assess quality of the included studies in these systematic reviews


Total unique citations on TB
[from PubMed & Embase for 2007-2008]
6459
Citations excluded:
No abstract available
1021 (15.8%)
Abstracts available
5438 (84.2%)
Abstracts excluded:
Not original study
215 (4.0%)
Case reports or case series
957 (17.6%)
Original Study
4266 (78.4%)
Coded as per UK Health Research Classification System (HRCS)
1. Underpinning Research
619 (14.5%)
1.1 Biological
545 (88.8%)
1.2 Psycological, Socioeconomic
1 (0.1%)
1.3 Chemical, Physical
9 (1.5%)
1.4 Methodology
52 (8.5%)
1.5 Ressources
7 (1.1%)
2. Aetiology
1607 (37.7%)
2.1 Biological
482 (30.1%)
2.2 Environmental
259 (16.2%)
2.3 Psychological, social, economic
92 (5.7%)
2.4 Distribution
708 (44.2%)
2.5 Methodology
60 (3.7%)
2.6 Resources
2 (0.1%)
3. Prevention of Disease and Conditions, and Promotion of Well-Being
209 (4.9%)
3.1 Behaviour
1 (0.5%)
3.2 Environmental
14 (6.7%)
3.3 Nutrition, Chemoprevention
0 (0%)
3.4 Vaccines
194 (92.8%)
3.5 Resources
0 (0%)
4. Detection, Screening and Diagnosis
699 (16.4%)
4.1 Preclinical
84 (12.0%)
4.2 Evaluation
584 (83.7%)
4.3 Impact
6 (0.9%)
4.4 Population Screening
12 (1.7%)
4.5 Resources
3 (0.4%)
4.6 Cost
9 (1.3%)
5. Development of Treatments and Therapeutic Interventions
328 (7.7%)
5.1 Pharmaceutical
316 (96.4%)
5.2 Cellular/gene
3 (0.9%)
5.3 Medical device
0 (0%)
5.4 Surgery
4 (1.2%)
5.5 Radiotherapy
1 (0.3%)
5.6 Psychological Behavioural
0 (0%)
5.7 Physical
0 (0%)
5.8 Complementary
2 (0.6%)
5.9 Resources
2 (0.6%)
6. Evaluation of Treatments and Therapeutic Interventions
279 (6.5%)
6.1 Pharmaceutical
189 (68.2%)
6.2 Cellular/gene
1 (0.4%)
6.3 Medical device
8 (2.9%)
6.4 Surgery
66 (23.8%)
6.5 Radiotherapy
3 (1.1%)
6.6 Psychological Behavioural
1 (0.4%)
6.7 Physical
0 (0%)
6.8 Complementary
5 (1.8%)
6.9 Resources
4 (1.4%)
7. Management Of Diseases and Conditions
349 (8.2%)
7.1 Individual care needs
155 (44.4%)
7.2 End of life Care
1 (0.3%)
7.3 Management
191 (54.7%)
7.4 Resources
2 (0.6%)
8. Health and Social Care Services Research
176 (4.1%)
8.1 Organisation Delivery of Services
104 (59.4%)
8.2 Health, welfare Economics
19 (10.9%)
8.3 Policy, ethics Research Governance
43 (24.6%)
8.4 Methodology
9 (5.1%)
8.5 Resources
0 (0%)

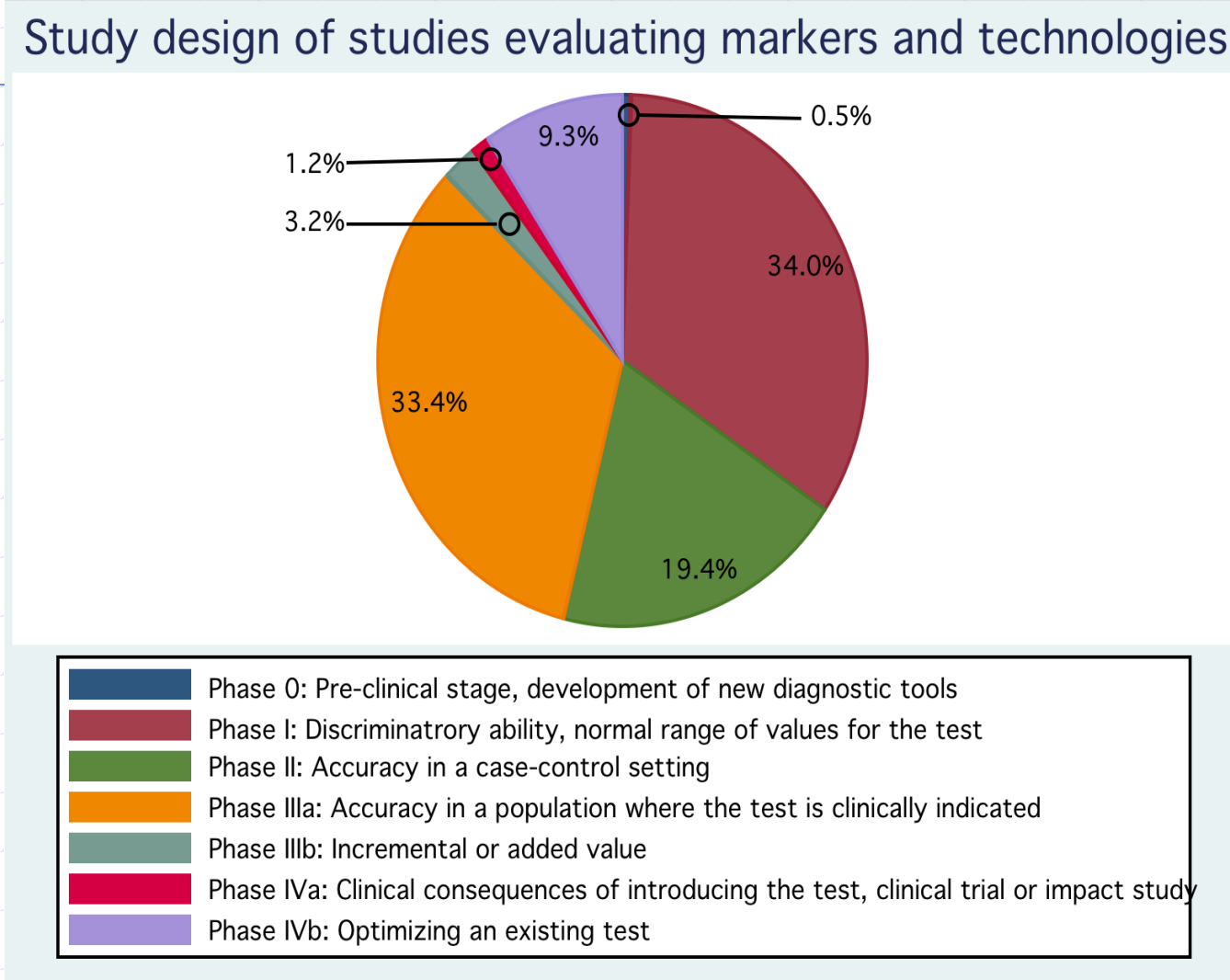
Distribution of major types of TB research activities [N=6459]



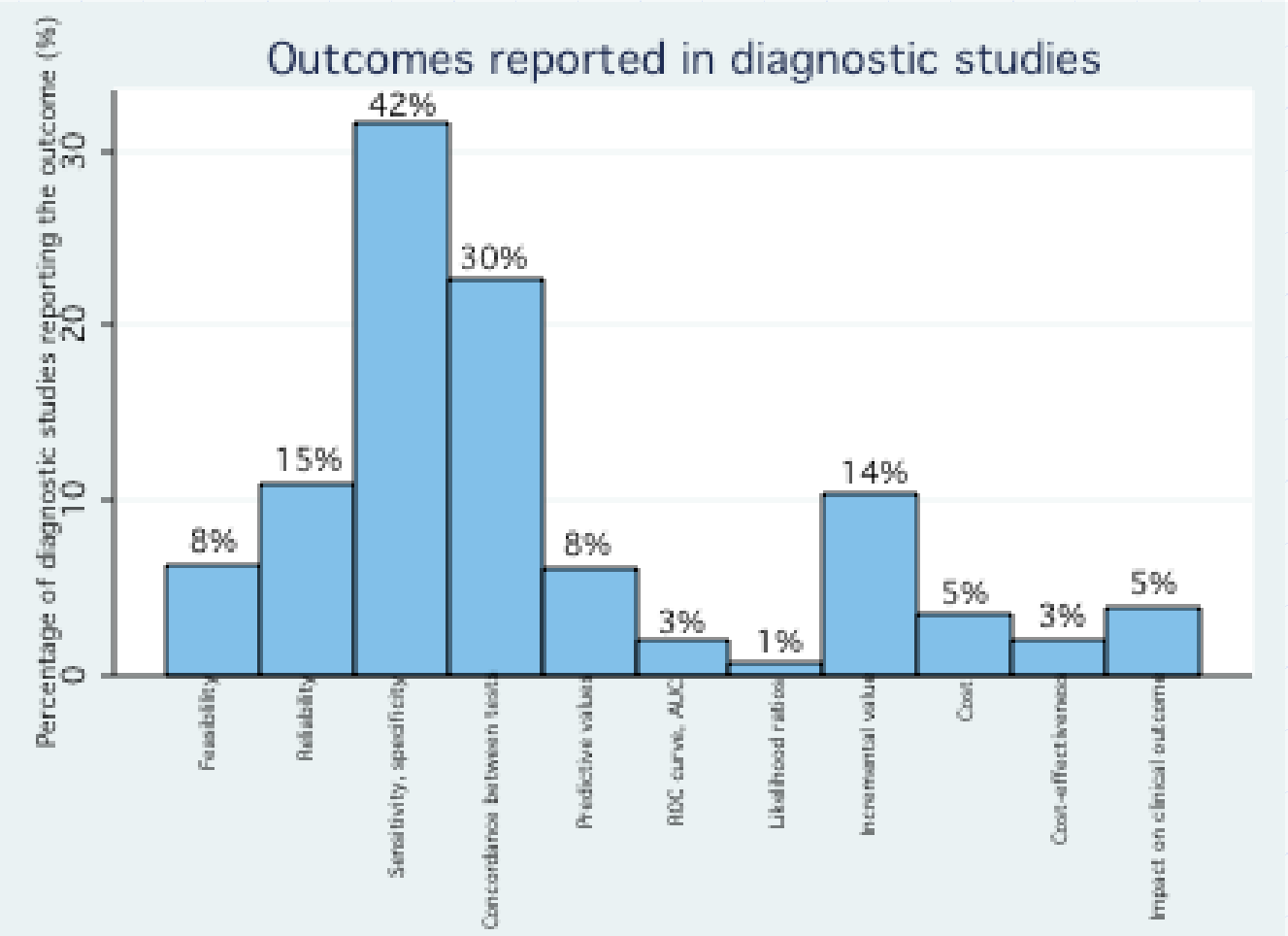
Distribution of study types within diagnostic research [N=**699**]



Distribution of phases within evaluation studies of diagnostics [N=584]



Distribution of outcomes reported in abstracts of diagnostic studies [N=**699**]



Results: quality and reporting of diagnostic accuracy 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]*

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.

Quality of TB accuracy studies using QUADAS [N=45]

Quality item	45 studies n (%)
Adequate spectrum composition	26 (58)
Clear description of selection criteria	21 (47)
Adequate reference standard	44 (98)
Absence of disease progression bias	42 (93)
Absence of partial verification bias	44 (98)
Absence of differential verification bias	42 (93)
Absence of incorporation bias	45 (100)
Absence of index test review bias	6 (13)
Absence of reference test review bias	7 (16)
Absence of clinical review bias	14 (31)
Report of uninterpretable results	9 (20)
Description of withdrawals	3 (7)

Fontela et al. PLoS One 2009

17 meta-analysis with over 500 diagnostic studies

- 52% (range 16 – 100%) of the trials used a prospective data collection design.
- 30% (range 0 – 95%) of the trials used a consecutive or random sampling method to recruit subjects.
- 75% (range 43 – 100%) of the trials used a cross-sectional design, and the case-control approach was used in about 25% of the studies.
- Any form of blinding was used in only 35% (range 0 – 78%) of the trials.
- In most studies (87%; range 10 – 100%), the index test results were verified by a reference standard test.

Table 2. Methodological quality of studies on tuberculosis diagnostics in recently published meta-analyses.

Meta-analysis	No. of studies	Diagnostic test	Average size of each study	Prospective data collection (%)	Consecutive or random sampling of subjects (%)	Cross-sectional design (%)	Blinded interpretation of test results* (%)	Complete verification of index test results‡ (%)	Ref.
Sarmiento et al. (2003)	16	PCR on respiratory specimens for smear-negative pulmonary TB	NR	50	NR	NR	63	100	[12]
Goto et al. (2003)	40	ADA for TB pleural effusion	137	NR	NR	NR	0	NR	[13]
Pai et al. (2003)	49	NAT for TB meningitis	42	61	49	61	59	94	[14]
Greco et al. (2003)	44	ADA and IFN-γ tests for TB pleural effusion	135	NR	NR	NR	9	NR	[15]
Pai et al. (2004)	40	NAT for TB pleural effusion	60	63	53	70	55	100	[16]
Flores et al. (2005)	84	In-house PCR for pulmonary TB	149	NR	NR	71	34	NR	[17]
Kalantri et al. (2005)	13	Phage amplification tests for pulmonary TB	448	NR	NR	85	23	100	[18]
Pai et al. (2005)	21	Phage-based tests for rifampin resistance	85	NR	38	NR	57	100	[19]
Morgan et al. (2005)	15	Line probe assay for rifampin resistance	91	NR	0	NR	13	100	[20]
Greco et al. (2006)	63	Commercial NAT for pulmonary TB	410	16	32	NR	16	NR	[21]
Steingart et al. (2006)	45	Fluorescence versus conventional sputum smear microscopy for pulmonary TB	493	100	36	NR	49	NR	[22]
Steingart et al. (2006)	83	Direct versus concentrated sputum smear microscopy for pulmonary TB	256	100	21	NR	31	NR	[23]

*At least single blind. ‡By reference standard.
ADA: Adenosine deaminase; IFN: Interferon; NAT: Nucleic acid amplification test; NR: Not reported; TB: Tuberculosis.

Pai M, O'Brien R. Exp Rev Mol Diagn 2006.

Conclusions

- 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

WHO policy process

- According to WHO, in order to consider a global policy change, WHO must have solid evidence, including clinical trials or field evaluations in high TB prevalence settings.
- Policy process includes a comprehensive review of the evidence, as well as expert opinion and judgment
- All WHO guidelines will be approved by a Guideline Review Committee
- All guidelines and policies will explicitly incorporate evidence using the GRADE approach

Box 1. WHO Policy Process for Tuberculosis

1. Identifying the Need for a Policy Change

The need to formulate new or revised policies may arise from WHO's ongoing monitoring of technical developments or from interested parties submitting requests with supporting documentation for policy or guideline development. WHO receives information about a new technology or approach via many channels, with the most direct lines coming from national TB programs and researchers themselves. To consider a global policy change, WHO must have solid evidence, including clinical trials or field evaluations in high TB prevalence settings.

2. Reviewing the Evidence

WHO may carry out or commission a review of the documentation of the technology's clinical or programmatic performance, including newly published and "grey" research or reviews, "proof of principle" reports, large-scale field trials, and demonstration projects in different resource settings. Standardized evaluation criteria have been and are being developed by the New Diagnostics, New Drugs, and New Vaccines Working Groups of the Stop TB Partnership.

3. Convening an Expert Panel

If the evidence base is compelling, WHO will convene an external panel of experts, excluding all original principal investigators from the studies. The panel will review the evidence and make a recommendation or propose draft policies or guidelines to WHO's Strategic and Technical Advisory Group for Tuberculosis (STAG-TB).

4. Assessing Draft Policies and Guidelines

STAG-TB provides objective, ongoing technical and strategic advice to WHO on TB care and control. STAG-TB's objectives are to provide the Director-General, through the Stop TB Department, with an independent evaluation of the strategic, scientific, and technical aspects of WHO's TB activities; review progress and challenges in WHO's TB-related core functions; review and make recommendations on committees and working groups; and make recommendations on WHO's TB activity priorities. STAG-TB reviews the policy drafts and supporting documentation during its annual meeting. STAG-TB may endorse the policy recommendation with or without revisions, request additional information and re-review the evidence in subsequent years, or reject the recommendation.

5. Formulating and Disseminating Policy

New WHO policies and guidelines will be disseminated through different channels to Member States, including through the World Health Assembly, WHO Web site, listservs, and journal publications. WHO also disseminates its recommendations to other agencies and donors engaged in TB control activities.

Source: World Health Organization [7]

http://www.who.int/tb/dots/laboratory/policy/en/index4.html

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

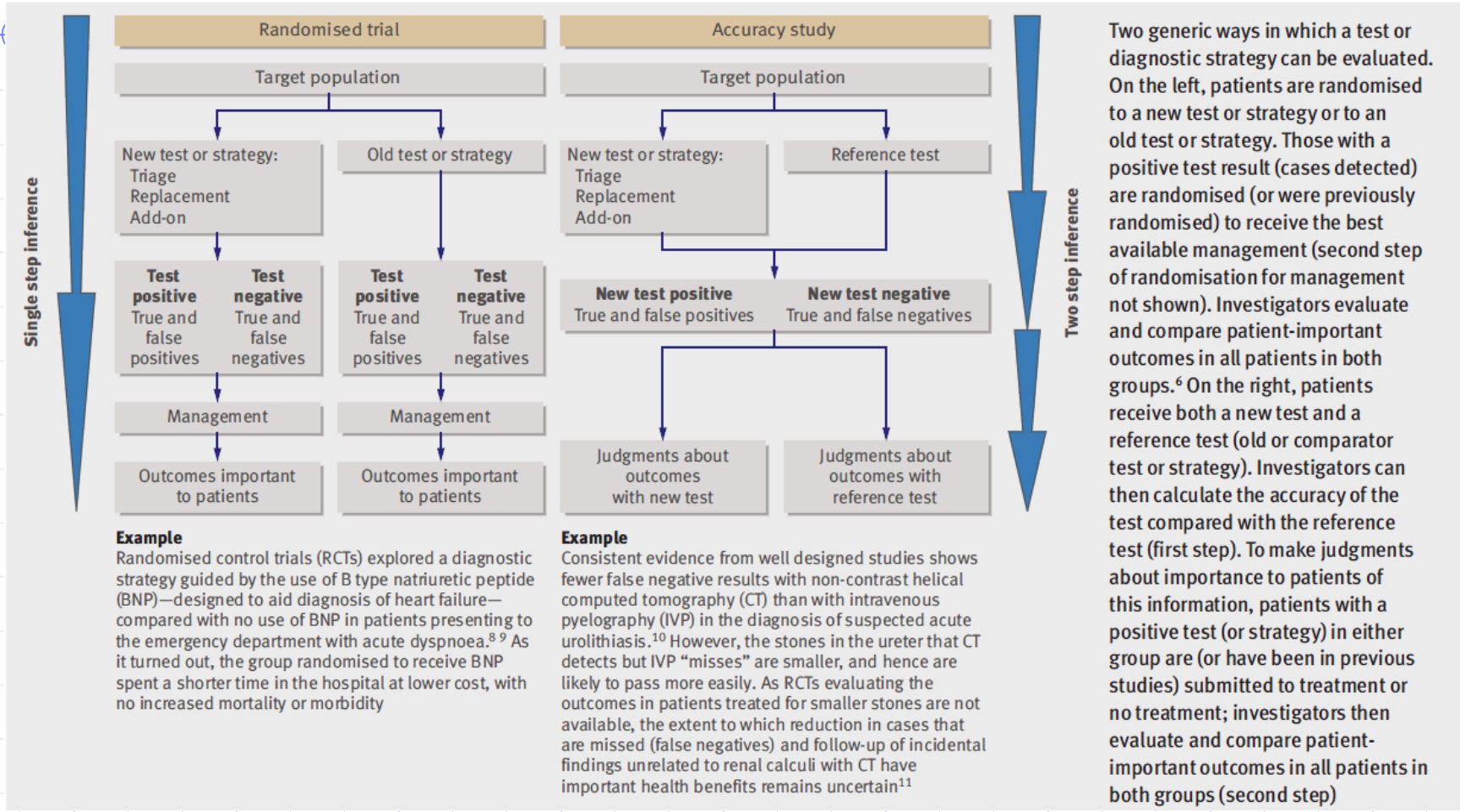
However, test accuracy is a surrogate for patient-important outcomes, so such studies often provide low quality evidence for recommendations about diagnostic tests, even when the studies do not have serious limitations

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

BMJ 2008

GRADE: for high quality evidence, impact on patient-important outcomes needs to be demonstrated



Schunemann H et al. BMJ 2008

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

- Example: Rapid diagnostics tests (RIDTs) for influenza
  - 100+ accuracy studies
  - 20+ impact studies (including several diagnostic RCTs)

In TB, since we have mostly accuracy data:

example from WHO EGM on tests for drug-resistant TB

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

There are 50+ systematic reviews on TB tests, but almost all focus on sensitivity and specificity (accuracy)

Evidence-Based Tuberculosis Diagnosis

A comprehensive resource for evidence syntheses, policies, guidelines and research agendas on TB diagnostics

- Home
- About NDWG
- What is E-B TB Dx?
- TB Diagnostics Pipeline
- Systematic Reviews
- WHO Policies
- Guidelines for TB Dx
- Research Agendas
- Resource Centre
- Diagnostic Research
- Related Links
- Credits
- Contact
- Events
- Latest News

Google™ Site Search

Search

Share

Evidence-based TB Diagnosis

www.tbevidence.org

Developed with the support of:

Stop TB Partnership's New Diagnostics Working Group (NDWG)
World Health Organization (WHO)
Foundation for Innovative New Diagnostics (FIND)
Special Programme for Research and Training in Tropical Diseases (TDR)
Global Laboratory Initiative (GLI)
Public Health Agency of Canada (PHAC)
Francis J. Curry National Tuberculosis Center, UCSF
McGill TB Research Group

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

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